DOCUMENT RESUME

ED 196 935 TM 810 050

AUTHOR Anderson, Beverly L.: And Others

TITLE Educational Testing Facts and Issues: A Layperson's Guide to Testing in the Schools.

INSTITUTION California State Dept. of Education, Sacramento. Office of Program Evaluation and Research.; Nero and Associates, Inc., Portland, Oreg.; Northwest Regional Educational Lab., Portland, Oreg.

SPONS AGENCY National Inst. of Education (ED), Washington, D.C.
PUB DATE Sep 80
CONTRACT 400-79-0059
NOTE 56p.; For related documents, see TM 810 047-049.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Educational Practices; *Educational Testing; Elementary Secondary Education; Lay People; Public Schools; *Testing Problems; *Test Interpretation
IDENTIFIERS *Test Use

ABSTRACT

This booklet addresses the role of testing in today's public education system, and presents a series of questions and answers which will be of particular interest to school board members, legislators, lawyers and journalists. These questions are grouped into two major categories: (1) test purposes and users; and (2) current testing issues. Current testing issues include how teachers view testing, why achievement test scores are declining, the meaning of the truth in testing legislation, the meaning of test bias, issues related to IQ testing, educational and legal issues surrounding minimum competency testing, and the evaluation of teachers in the schools. In addition, an annotated bibliography, a glossary of measurement terms, and a summary of common test scores are included that aid in the layperson's quest for information related to the issues. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Educational Testing Facts and Issues:

a layperson's guide to testing in the schools

Beverly L. Anderson
Richard J. Stiggins
David W. Gordon

National Institute of Education, U.S. Education Department
Contract No. 400-79-0059

Coordinated by:
Nero and Associates, Inc.
520 S.W. Sixth Avenue, Suite 820
Portland, OR 97204
Susan W. Rath, Project Director

Materials developed by:
Northwest Regional Educational Laboratory
Assessment and Measurement Program
710 S.W. Second Avenue
Portland, OR 97204

California State Department of Education
Office of Program Evaluation and Research
721 Capitol Mall, 4th floor
Sacramento, CA 95814

Acknowledgements:

Special thanks are due to the many workshop participants and sponsors who provided helpful comments during the development of this booklet. Appreciation is also due to the legislators, school board members, journalists, measurement specialists, lawyers, and test publisher representatives who reviewed it, and Carol DeWitte who was responsible for its production.

Designed and illustrated by Warren Schlegel

Edited by Jane Loftus

September 1980

This booklet is intended to be used in conjunction with workshops and seminars conducted by measurement specialists using the training methods described in Training Citizen Groups on Educational Testing Issues: A Trainer's Manual, developed under this same contract.

These materials are in the public domain and may be reproduced without permission. The following acknowledgement is requested on materials which are reproduced: Developed by the Northwest Regional Educational Laboratory, Portland, Oregon and the California Department of Education.

This booklet was prepared by the Northwest Regional Educational Laboratory, a private nonprofit corporation and the California Department of Education under a subcontract with Nero and Associates, Inc., Portland, OR. The work contained herein has been developed under a contract with the National Institute of Education, U.S. Education Department pursuant to Contract No. 400-79-0059/SB0408(a)-79-C-197. The opinions expressed in this publication do not necessarily reflect the position of the National Institute of Education, and no official endorsement by the Institute should be inferred. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

Table of Contents

	Page
INTRODUCTION	1
OVERVIEW OF TEST PURPOSES AND USERS	3
Who uses tests?	3
What are the most common types of tests?	3
What are the major purposes of testing?	4
What are limitations of tests?	7
Who is responsible for initiating testing?	8
Who constructs tests?	9
What are the costs of testing?	9
CURRENT TESTING ISSUES	11
How do teachers view testing?	11
Why are achievement test scores declining?	11
What is the meaning of the truth in testing legislation?	13
What is the meaning of test bias?	14
What are the issues related to IQ testing?	17
What are the educational and legal issues surrounding minimum competency testing?	21
Are tests being used to evaluate teachers in schools?	26
ANNOTATED BIBLIOGRAPHY	29
APPENDIX A: A GLOSSARY OF MEASUREMENT TERMS	35
APPENDIX B: SUMMARY OF COMMON TEST SCORES	45

Introduction

This booklet addresses the role of testing in today's public education system, and presents a series of questions and answers which will be of particular interest to school board members, legislators, lawyers and journalists. These questions are grouped into two major categories:

- Test Purposes and Users
- Current Testing Issues

Before presenting these issues, a short scenario from a typical school may help in establishing a context for the role of testing in schools today.

An interviewer recently visited a junior high school to learn more about the role of testing in the school. Walking down the hall, the first person the interviewer met was a student leaving a room marked with a sign "Testing - Do Not Disturb."

The interviewer said, "Hi! I'm visiting your school and want to find out what kind of testing is done here. It looks like you just took some tests."

"Yes," the student replied. "We're taking a series of tests this week to find out what classes we should be taking. They just gave me some tests in math and reading."

The interviewer asked a teacher about the testing that was being done. "Yes, we use those results to group students. But if a teacher disagrees with the placement of a student, the teacher's opinion is taken into account as well as the test results."

After several more stops, the interviewer found that in the history and social studies classes, no standardized achievement tests were given; rather, all the testing done in those classes was designed by the classroom teacher.

At the district testing specialist's office located at the junior high, the interviewer discussed the district testing program with the specialist.

INTERVIEWER: What are the major reasons for testing in your district?

SPECIALIST: The districtwide testing is for three major purposes: first, to determine trends in student performance over the years; second, for program evaluation; and third, to determine student placement. Diagnostic testing is done at the discretion of teachers and principals. It is not determined at a district level.

INTERVIEWER: What types of tests are used?

SPECIALIST: Let me give you an example of what a typical student would experience in grades K through 12. During their first two months in kindergarten, students are given a screening test. It is essentially an observation of a student's physical development, verbal and other academic skills.

In grades 1 through 6, the student takes a standardized reading and math test each spring. In grades 7, 9 and 11, the student takes a language arts test as well as the reading and math test. In grades 3, 7, 9 and 11, an aptitude test is given along with the achievement battery. The purpose of the aptitude test is to establish expected levels of performance on the achievement test.

INTERVIEWER: How many hours of testing do you think the typical student experiences?

SPECIALIST: Well, the districtwide testing I mentioned takes about two hours in the first grade with the amount of time increasing progressively to nearly six hours in the fifth grade. From the fifth grade on, it fluctuates between four and six hours.

INTERVIEWER: What about students who are having difficulties in certain areas or appear to be in need of special education?

SPECIALIST: Now you have hit on an important purpose for testing. Students in special programs such as Title I, Follow Through, or a bilingual program experience much more testing. Nearly all federal or state funded programs require program evaluation; typically, students are tested both in fall and spring for this purpose. We wish the testing could be coordinated with districtwide testing, but an evaluation frequently requires a different test; thus these students take at least two more tests during the year. Furthermore, programs like Title I frequently require diagnostic testing throughout the year. Students in such programs may participate in double or triple the amount of testing of the typical student.

INTERVIEWER: I hear a lot about minimum competency testing. Are you doing such testing in your district?

SPECIALIST: Not yet, but we will be starting next year. Our school board feels that minimum competency testing will be very useful in identifying students who should receive remedial instruction. They are still debating whether or not to require passage of the test for graduation. They have decided to wait until after next year's testing to decide. We have spent a lot of time this year working with teachers, administrators and community members to decide what competencies to test with the MCT, as we call it. We contracted with an educational service agency to prepare the test once we had the competencies and skills identified.

INTERVIEWER: Are people concerned about cultural bias in testing?

SPECIALIST: Yes, there is much talk about cultural bias. Unfortunately there are so many different interpretations of what cultural bias is that we have a very difficult time dealing with it. I'm going to a workshop next month on the topic which will hopefully help me determine how to handle this issue. Partly out of concern about cultural bias, we are seriously considering eliminating our aptitude testing, but I'm not ready to recommend that yet.

INTERVIEWER: Another topic I am hearing more and more about is teacher evaluation and the use of testing for that purpose. Is that an issue in your district?

SPECIALIST: Do you mean the use of student test scores in evaluating teacher performance or actually testing teacher competencies?

INTERVIEWER: I was thinking of the former but both topics are of interest.

SPECIALIST: Because of the many problems inherent in using student test scores for teacher evaluation, we do not use them for that purpose. We are getting pressure from parents, however, to at least consider looking at the scores of students over several years when a particular teacher's performance is questioned. As far as testing teachers, we just started giving teacher applicants a test of basic skills competencies. Teachers already in the district are not tested.

The district described in this imaginary interview is meant to be representative of many districts across the country. The issues raised here are discussed in the following pages.

Overview of Test Purposes and Users

Who uses tests?

Tests are used by many people. Teachers use tests to determine students' progress in learning specific skills. Parents use test scores to tell them how their child is doing in school or to see how their school compares with other schools. School board members and legislators use test data to help set policy and allocate funds. School principals, guidance counselors, district personnel and state department of education staff also require informatin on how well students are learning. News reporters often request student test scores for reports on quality of schools. Lawyers may find test scores to be important in certain legal cases. State, federal, or private agencies which fund special programs often require student test scores to evaluate the program's effectiveness. And, of course, students use test scores to determine if they are learning what they are expected to learn.

What are the most common types of tests?*

There are several types of measurement devices used in the schools. Some tests measure knowledge and skills and some measure other characteristics. There are two main types of cognitive measures used in today's elementary and secondary schools--achievement tests and aptitude tests. Other measures such as attitude inventories and interest inventories are also used.

ACHIEVEMENT TESTS

These tests measure how much a student has learned or what skills the student has acquired. Achievement tests are developed by teachers for classroom use or by test publishers for use by schools and school districts in large-scale testing programs. In either case, the test is developed by outlining the material to be tested and writing test items representative of that material. Achievement test scores are used by teachers and students to help plan and manage instruction (diagnose weaknesses, assign grades, etc.), to certify mastery of minimum essential skills, to select students for admission to college, to plan career directions, and to evaluate the quality of educational programs.

Achievement tests come in two basic forms: those used to compare one student's learning with that of another student and those used to determine if a student has mastered particular knowledge and skills regardless of how other students score. Many achievement tests given are standardized. These tests cover material taught in most schools in subject matter areas such as reading, language arts, mathematics, science, and social studies. Once developed, the tests are administered to large national samples of several thousand students. Student performance is then analyzed and ranking by scores is established. These comparative or norm referenced tests are then used at the local district level where they allow the comparison of student test scores within the district. For

*See Appendices for a glossary of measurement terms and descriptions of test scores.

example, a student may be at the 40th percentile compared to a national norm group, but at the 50th percentile compared to a local norm group. This would indicate that the district as a whole was performing lower than the national group.

Norm referenced tests are used to select students for remedial or advanced programs. In addition, these tests are used as a guidance tool for the long-term educational and vocational planning of the student.

Achievement tests can also show the quantity of specific knowledge and skills (learning objectives) that the student has mastered. These tests, known as criterion or objective referenced tests, are most useful for diagnosing specific strengths and weaknesses in individual students, for certifying mastery of minimal competencies, and for evaluating specific educational programs. Objective referenced tests are most often developed by teachers. However, nearly all major test publishers have objective referenced tests available. In some cases, test publishers may provide both objective and norm referenced interpretations for the same test. Increasing numbers of local districts employ testing specialists to develop their own objective referenced diagnostic tests--either for districtwide testing or for local diagnostic use by teachers. Some states, California, Michigan, Oregon, Texas and New Jersey, among others, are also developing objective referenced tests for statewide assessment purposes.

APTITUDE TESTS

Aptitude tests are designed to measure the ability to do school work. These tests can measure the ability to use language, to solve problems, to deal with mechanics and to think in terms of mathematics. These abilities are not inherent or unchanging. They can be influenced by many factors: experience, family, culture, emotions and health. Aptitude relates to achievement in that abilities provide a basis for achieving. Aptitude influences the amount of learning that takes place. Aptitude test scores are commonly norm referenced or comparative.

A summary of the various test scores commonly used for the different cognitive measures is presented in Appendix B.

ATTITUDE INVENTORIES

Another common test investigates how students feel toward school, or toward a particular subject or person within the educational system. Such inventories are frequently used in evaluating special programs. Seldom are they administered districtwide. While such measures are available from commercial publishers, these inventories are usually developed locally to answer questions of interest to a particular district. They often have low or unknown validity and even when appropriately used must be interpreted cautiously and in conjunction with other data.

INTEREST INVENTORIES

These instruments attempt to pinpoint any interest that may influence a student's learning or career plans. Usually a guidance counselor or teacher has responsibility for interpreting the results.

What are the major purposes of testing?

Tests are used for three purposes: instructional management, entry-exit decisions and programmatic decisions. Instructional management

and entry-exit decisions require test data for each student. Programming decisions can be made based on group data, which allows a sampling of students rather than testing every student.

INSTRUCTIONAL MANAGEMENT

Tests play an important role in instructional management decisions. Data from these tests are used for the diagnosis of students' strengths and weaknesses, student placement, and educational-vocational student guidance.

Diagnosis. Perhaps the most frequent use of tests is to diagnose the educational development of individual students. Here, the teacher is the primary decision maker, although students may also be involved. Teachers often use tests and other performance indicators to assess the student's current development so that the next, most appropriate instructional unit is selected. Tests useful in diagnostic decision making are those that reveal precisely what skills and knowledge the student has or has not mastered.

Placement. If diagnosis determines what instructional units within a course a student needs to master, then placement groups the student according to the next level of instruction best suited to that student's skills. In this case, the decisions are made by administrators, teachers, and guidance counselors who must place each student in the most appropriate course. Math tests, for example, might be used to place students at the appropriate level in a high school math course sequence. A test which indicates student ability in math will ensure that students will not be assigned to courses which are too advanced or too elementary for them. Placement tests usually cover a broader range of knowledge and skills than diagnostic tests and are only used once or twice a year. Diagnostic tests may be used on a day-to-day basis. However, completion of grades and courses are also considered in placement decisions.

Testing is the major method used to identify students who would benefit from placement in special programs (bilingual programs, special education programs, remedial reading and math) or particular educational experiences. Standardized achievement tests are the most frequently used measures for placing students in compensatory education programs. In addition, aptitude and psychomotor tests are often used to identify students who need special education.

Guidance. While diagnosis matches the student to an instructional unit, and placement matches a student to a course, guidance can determine an entire program of study. Here, students and their parents assisted by guidance counselors make the decisions. When students decide which educational and vocational program to pursue, they must consider their chances of success and satisfaction. These career planning decisions, typically made in junior and senior high school are assisted by the use of tests that cover broad academic areas and tell the students where they stand in relation to other students. These tests scores can also determine students' strengths and weaknesses which will aid them in making choices. Test scores, of course, should never serve as the sole basis for any guidance decision. The student's academic record, interests and aspirations all merit consideration.

Guidance testing, which is generally determined by school or district administrators and guidance counselors, is usually a secondary result of placement or diagnostic testing.

ENTRY OR EXIT DECISIONS

Tests are also used to determine if a student should be placed in an educational program or to determine if a student has completed a program's requirements. For example, tests may be administered in order to select students for programs with limited enrollment (e.g., college entrance or trade school), or to certify minimum competencies (e.g., for high school graduation or occupational licensing).

Selection. The difference between selection and placement is not always clear. Placement, as previously described, groups students in the most appropriate level of instruction. This is an instructional management decision. Selection refers to a process whereby students are screened for admission to an educational program which has a limited number of participants. Admission is based on who is likely to benefit. Here, the key decision makers are teachers and administrators. A test used for the purpose of selection focuses on students' skills and knowledge considered essential for success in the program, and compares students' relevant skills and knowledge so that those most likely to succeed are identified. Admission to college or into a particular course (for example, airline pilot training) are prime examples of selection. However, test scores are not the sole basis for selection decisions. Previous academic record and other performance criteria may also be considered.

Perhaps, the most common use of selection testing is the college entrance examination. Colleges require a specific entrance examination and interested students register with test publishers who carefully control the administration of the tests at various locations across the country.

Certification. Tests often play an important role in certifying acceptable minimum levels of educational development in students. For example, a teacher might use a test to certify mastery of beginning verbal skills required for completion of a certain course. Or, a district administrator applying Board of Education graduation standards might use an examination in order to test a student's mastery of minimally acceptable skills. Or, members of a certain technical profession might use a test to certify competence in that profession. Since, in each case, those taking the exam must pass the test to be certified, the test must focus specifically on clearly stated minimal competencies.

PROGRAMMATIC DECISIONS

A third use of tests is to assist in program planning. In this instance, test data may be helpful in providing the basis for developing a new program, allocating funds or evaluating existing programs. Such testing falls into three categories: survey assessment, formative program evaluation and summative program evaluation.

Survey Assessment. Probably the most common use of testing in education is to survey student achievement and analyze trends over time in order to assist in program planning. This kind of testing is usually designed to raise issues for further investigation. For example, the test results might prompt such questions as, why are math scores gradually declining in the district (or state or nation)? Or, why are reading scores of fourth graders consistently below national averages while those in other grades are above average? The test data are used to identify which aspects of the educational system need to be more thoroughly investigated as well as possible reasons for unsatisfactory performance. For this purpose, achievement test scores--sometimes

from random samples of students--are gathered annually, then averaged across the entire school, district or state, and used to indicate the level of student development. In order to show trends, test scores are frequently compared from year to year. This information then becomes a basis for setting educational policy and allocating funds. Typically, educational administrators are the primary decision makers, but they must justify these decisions to the ultimate decision maker, the taxpayer. Tests used to assess an educational program must cover broad content and skill areas in order to provide valid information for program changes.

Formative Evaluation. In formative evaluation, the goal is to determine which instructional units or features of a specific educational program (e.g., remedial reading), are effective and which need revision. In this instance, tests are used to measure what the students learn in a specific program and the results are used to help shape or revise the program during its formative stages.

Summative Evaluation. Summative evaluation reveals a program's overall merit, and suggests whether or not a program should be continued, terminated, or expanded. Tests designed to assess knowledge gained from a program are an important part of such an evaluation. Teachers, program, building or district administrators, and the public, represented by the board of education, may be involved in summative evaluation decisions. Tests may be given both before and after instruction, with retesting after an interval to determine the student's retention of knowledge.

It should now be obvious that tests are used for many different purposes in education. Many decisions using test data affect individual students, while other decisions affect whole groups. The implications of these decisions vary. Some have far-reaching, long-term effects. Others are less serious. Tests can be valuable, but test selection must be made carefully.

What are the limitations of tests?

Test users should consider that tests represent only one of many types of performance indicators. In the classroom, day-to-day classroom activities and classwork represent important and valuable sources of information about student development that should be used to supplement test information in making educational decisions. Tests are also supplemented with professional teacher judgments.

Tests are designed for certain uses; a single test cannot serve all purposes. Tests are limited in terms of the range of decisions they can help with. Generally, a test is capable of assisting in one or two of the decisions previously discussed. The key to using tests effectively is to know what decision is to be made, to determine what material needs to be tested to aid that decision and to be certain that the test used actually covers that material.

Tests are also limited in the material they cover. Generally, tests cover only a sample of the content or skills taught. It is almost never feasible, both in terms of time and money, to test every aspect of the subject matter taught. As a result of this sampling procedure, as well as uncontrollable factors such as motivation and fatigue, test scores are subject to some variability. That is, if the same test was taken twice by the same student, the score might vary slightly due to the imprecision of the test. Therefore, a score should seldom be seen as completely precise or unchanging. Rather, it should be seen as a general performance index.

Another limitation of tests is that they are easy to misuse. They are readily available and relatively easy to construct, especially if quality is disregarded. Therefore, they are easy to misuse. Misuse can only be avoided by knowing precisely how the test score is to be used and by selecting or building a test specifically designed to serve that purpose.

Who is responsible for initiating testing?

Often it is assumed that tests are initiated, for the most part, by teachers who need information to improve instruction. This is generally true, however, mainly of teacher-made tests and curriculum-related tests. It is not the case with most standardized tests or district and state-developed tests. Decision makers at all levels--federal, state, and district--need information from these tests.

At the federal level, the primary impetus for testing comes from federally-funded special programs, which usually require the evaluation obtained by using standardized achievement testing. Title I of the Elementary and Secondary Education Act, which provides funding for compensatory education, is a case in point. As the largest single item in the United States education budget, Title I programs are subjected to rigorous evaluation to demonstrate effectiveness. Although current Title I evaluation procedures require local programs to either use standardized tests or the combination of nonnormed tests and a standardized test, specific recommendations for which particular tests to use are carefully avoided.

At the state level, the most common reasons for testing are statewide assessment for accountability, minimum competency, and for evaluation of state-funded special programs. Legislators, who wish evidence that schools are doing the job they're being funded to do, often call for statewide assessment testing. The late 1960's saw many such assessment programs established. Following the state assessment movement was the public outcry for students to achieve certain minimum competencies before high school graduation. In response, at least 38 states have enacted legislation requiring minimum competency testing. Evaluation of state-funded special programs also provides an impetus for state-level testing.

Generally, federal and state regulations allow state and local education agencies considerable latitude in setting their own testing procedures. For example, although Title I evaluation requires the use of standardized tests, many different standardized tests are available. Although states may put some limitations on which tests are acceptable, final selection is generally a local decision.

Most district-initiated testing is done to ensure accountability, to place students in special programs, to evaluate program results, and make instructional management decisions. Typically, the district decides which test is to be used for evaluating federal- and state-funded special programs. District level testing policy beyond that required by federal and state regulations is determined by many factors: public pressure for accountability, teacher and administrator demands that tests be reflective of program goals and content, pressures from teachers' associations to avoid using student test results in teacher evaluation, and requests from teachers and administrators to reduce the amount of testing. District administrators and school boards are frequently in a quandary when establishing a testing

program that responds to these conflicting pressures. At the building level, the amount of additional testing beyond district requirements varies greatly. Generally, districts allow schools considerable autonomy, and the principal's perspective on testing can be a major influence.

At the classroom level, teachers as individuals or teams often conduct additional testing at their discretion. Some teachers employ comprehensive diagnostic systems, particularly in the basic skill areas of reading and math. They also may administer unit tests which accompany textbooks. Teachers generally need more diagnostic test information on lower performing students than on others.

In general, frequency of tests is determined by federal, state and district mandates for evaluation, accountability, student placement and certification rather than by requests from teachers or local administrators.

Who constructs tests?

Until recently, tests were almost exclusively constructed by either the classroom teacher or the commercial test publisher. But within the last 15 years, state departments of education and local school districts have begun to develop their own tests.

Classroom teachers generally construct tests to measure the specific instructional content being taught. These tests often take the form of a short weekly quiz, a mid-term examination or an end-of-the-course test. The test results are primarily used for grading or for helping students identify specific course content which they have not mastered.

The most frequently used tests developed by commercial publishers are the standardized achievement and aptitude measures. These tests require careful development of questions as well as extensive administration to establish interpretable test scores. During development, tests are administered to a carefully selected sample of students in a specified age or grade level. The results are used to establish scales which permit comparison of a student's score to national averages. The development of these "normative" scales is a costly process.

Commercial publishers also develop criterion or objective referenced tests. These tests are not tied to any one textbook series, but are focused on particular knowledge or skills that can be taught by a variety of methods or materials. These tests, for example, may measure a student's ability to add whole numbers regardless of the textbook or method of instruction used.

Publishers also develop tests which are contained in or related to specific textbooks. These tests, which may be used at the end of a unit, are tied to information in a particular text or set of curriculum materials.

The tests developed by state departments of education and local school districts are frequently designed to measure the school's success in teaching course content considered important in that state or district. Publishers' tests, based on the content most frequently taught across the nation, may not exactly match local curriculum content. Such tests should be carefully screened and selected to match local needs.

What are the costs of testing?

The actual cost of testing varies with the type of test used and its origin. For instance, objective tests scored by counting the number of test items answered correctly, and

performance tests which require the observation and evaluation of a process or product by a qualified judge differ in cost. These tests may be purchased from a test developer or test publisher, or they may be developed by local educators for local use. The costs of testing depend on the combination of these factors.

In all cases, there are three categories of costs: developmental costs, costs of test administration, and test scoring costs.

When an objective test is purchased, developmental costs include (1) the cost of time required to plan the testing context which includes thinking through the decision to be made and the kind of test needed, (2) the cost of time to review available tests, and (3) the costs incurred in actually purchasing test booklets, answer sheets, administration manuals, etc. Test administration costs will include time to (1) plan test administration, (2) train test administrators, (3) coordinate distribution of materials, and (4) administer the test and collect materials. Test scoring costs include (1) the time required to count the items answered correctly or (2) costs of optical scanning and computer scoring of answer sheets. There are also costs involved in disseminating the scores and interpretative information to the decision maker in a timely manner.

When an objective test is to be developed locally for local use, developmental costs include time required to (1) plan the test context, (2) write the test items, and (3) assemble the final test. If the test is to be used for very important large group decisions such as certifying proficiency for graduation, additional developmental costs will be incurred to pilot test the items before they are used in order to ensure a high quality test. Test administration and scoring costs will be the same as those previously discussed.

When a performance-based test is to be used, the scoring becomes more expensive because qualified judges must be used to score the test. When such a test is to be purchased, developmental costs include (1) time to plan the test context, (2) time to locate, review and evaluate available test exercises and scoring (rating) procedures, and (3) the costs of purchasing test materials. Test administration costs will generally be the same as those involved in the objective test. Test scoring costs, when such tests are used on a large scale, include time required to (1) plan scoring procedures, (2) select judges, (3) train judges, (4) score the test, and (5) process scores for the decision makers. Individual classroom use of these tests requires only planning the scoring procedures, scoring the test, and preparing results.

And finally, when a performance test is to be locally developed for local use, the test developer must (1) plan the test context, (2) develop exercises, (3) plan scoring standards and procedures and (4) conduct quality control research (for large-scale use). Test administration and scoring costs will be the same as those discussed above.

The point is that there are real and significant costs associated with sound (fair and useful) testing. However, money spent for good assessment will pay dividends in the form of high quality educational decisions.

Current Testing Issues

In view of the variety of test purposes and users previously discussed, there are several important issues that need to be addressed.

Issue 1: How do teachers view testing?

Throughout the educational community there is growing concern about the role of testing in the schools. At all levels - federal, state, and local - educators are aware of the possibility of overtesting. Administrators are reviewing testing programs to ensure that the fewest number of tests are being used and that the purposes for testing are clearly defined. Teachers as well as other educators are opposed to tests which damage a student's self-concept, perpetuate negative expectations, are biased against economically disadvantaged students or students with different cultural or linguistic backgrounds, or which are used as the basis for inappropriate comparisons of students or schools. Many educators are also opposed to the use of standardized tests for teacher evaluation and are particularly concerned that tests not be used as the sole criterion for important educational decisions. They are, however, supportive of testing to diagnose learning needs, prescribe instructional activities and measure progress in the curriculum content using tests prepared or selected by classroom teachers. Two major teachers' associations, the National Education Association and the American Federation of Teachers have taken steps to investigate the issue of testing. For example, the National Education Association last year published two booklets, Parents & Testing and Teachers & Testing (see bibliography) to assist its members in understanding testing issues. The American Federation of Teachers is in the process of preparing a handbook to improve understanding and use of standardized tests in the classroom.

Issue 2: Why are achievement test scores declining?

Since the mid-1960s there has been a well-publicized decline in the achievement test scores of students in the United States. This decline has been found in nearly all subjects and all regions of the country, and in almost all national testing programs, ranging from college entrance tests to elementary school achievement test batteries. Although precise amounts of score decline are difficult to determine, declines tend to be more pronounced through the higher grade levels and there seem to be differences in decline between male and female students. As we move into the 1980s, there is some evidence that the decline may have leveled out, but year to year test score patterns will have to be carefully observed in the future.

During the mid and late 1970s, a great deal of educational research focused on reasons for the decline. Early studies dealt with explanations related to test characteristics, hypothesizing that the decline might be a technical, rather than a real, phenomenon. These hypotheses were not supported[1], leading to the

[1]See Modu, C.C. and J. Stern. The stability of the SAT score scale. Research Bulletin RB-75-9, April 1975,. Educational Testing Service, Berkeley, CA.

conclusion that the decline was a real and significant socio-educational fact. Subsequent efforts focused on social-educational reasons for the decline.

One example is the work done at CEMREL, a research institute in St. Louis. In this study (consult annotated bibliography for complete reference), researchers collected and summarized evidence on the test score decline and sought possible causes in the school environment. Information was gathered and interpreted on the potential role of such factors as curriculum, course enrollments, and amount of schooling, as well as television watching and family background and environment. The researchers concluded that there is no evidence of changing teacher qualifications, and school organization and student motivation do not seem related to the decline. However, there is evidence of declining drop out rates accompanied by increasing absenteeism. This has the effect of leaving more low-achieving pupils in school. There is also evidence of a pronounced decline in the number of and enrollment in academic and college preparatory courses in high schools. In addition, some evidence was found that such non-school factors as TV watching, drug use, and family structure are potential contributors to the decline. From these initial exploratory efforts, the researchers concluded that there are many causes for the score decline and much added research is needed to provide a more concrete explanation for achievement drops.

Two additional attempts to find explanations for the declining college admission test scores were conducted by the College Entrance Examination Board (CEEB) and The American College Testing Program (ACT). CEEB formed an advisory panel of noted scholars and educators to examine the decline in Scholastic Aptitude Test (SAT) scores. After a year of study, the committee concluded that the decline can probably best be explained in terms of changes in the population of students taking this particular test and changes in the socio-educational fabric of the United States. Since SAT and ACT tests are taken by a select group of students, the panel concluded that the current SAT tested group is more broadly representative of American youth today than it was a decade ago when colleges were being more selective. Factors discovered to influence the socio-educational environment included increasing electives in high school, declining seriousness of educational purpose in society, television watching, changing family roles, the social unrest of the early 1970s, and motivation of students.

ACT assembled evidence of declining ACT Assessment Program test scores and combined it with evidence from other national testing programs to conclude, as had CEEB, that the college bound student population is changing. With more middle and low achieving students now considering college and participating in college-entrance testing--because of available opportunities and financial aid--the effect has resulted in a lowering of the average test score. In this instance, the test score decline could be interpreted as evidence of increasing diversity in educational opportunity--a positive statement--rather than an indictment of the educational system.

The conclusion from these studies is that there is no single explanation for the decline in test scores. Rather, a large number of complex factors has caused the score patterns we now observe. However, even in the absence of a clear explanation for the decline, the publicity it has received has had a pronounced impact on schools. That impact has been felt in testing and instruction. Teachers have carefully scrutinized the tests used to show declining achievement and

have challenged their appropriateness. And in response to the demand for alternatives, newly developed and specifically focused minimum competency tests covering relevant school and life skills have emerged. The effects on instruction have also been profound. Much more attention is being given to basic skills instruction in reading, writing and math from elementary school through college.

Issue 3. What is the meaning of the "Truth in Testing" legislation?

The debate over "truth in testing" resembles many of the arguments over consumer protection laws in the 1960s. At the center of the debate are two definitions of "fairness." On one side are the proponents of disclosure legislation, who argue that as a matter of simple fairness students should be able to see the test instrument (including the questions, the answers and related test data) used to make important decisions about their lives. Proponents feel that tests are social policy instruments that should, in a democratic society, be open to scrutiny. The opponents of such legislation argue that test security insures fairness, so disclosure of the tests will, by breaching security, affect the validity of the tests, increase the costs and lessen college admissions officers' confidence in standardized tests, all of which will make fair decision-making more difficult. They feel that secure standardized tests give everyone an equal chance and are more democratic instruments for policy making than are alternatives that permit the introduction of various biases.

Proponents of the legislation believe that the principle of fairness outweighs technical objections to open testing. They contend that security is not essential for test validity and that the burden of proof rests upon the test companies. Specifically, they ask that the test companies prove their allegations that full disclosure will weaken test validity, increase development costs, exhaust the number of test questions that can be asked, erode confidence in tests and lead to unfairness in decisions that involve test scores.

Opponents of the legislation, on the other hand, argue that the burden of proof rests upon the supporters of testing legislation. They ask for proof that the allegation that a substantial problem with test use or abuse exists, that the legislation will correct any misuses and abuses, that the added complexity of test development required for open testing is necessary and that substantial benefits will accrue to individuals and society through test disclosure.

CURRENT LEGISLATIVE ACTION

The first law requiring test publishers to disclose information to test takers and the public was California's SB 2005, enacted in September 1978. The law applies to any standardized test used for postsecondary education admissions selection of more than 3,000 students--in other words, such tests as the Scholastic Aptitude Test (SAT) and the American College Testing (ACT) Assessment. The law requires that a test's sponsor must file with the California Postsecondary Education Commission various kinds of data describing the test's features, limitations and use; must provide test takers with various kinds of information about the test and how it will be used; and must submit data about the administration of the test, the income realized and the expenses incurred in its administration.

New York enacted a similar law in 1979. Like the California law, it

applies only to tests used for postsecondary or professional school admissions and requires test publishers to file background reports about their tests and provide test takers with test information. In addition, the New York law requires the test agencies to file the contents of the tests with the New York Commissioner of Education within 30 days of release of scores, and, thereafter, to provide them to test takers upon request.

In addition to these laws, similar bills--some requiring total disclosure of the test (such as the New York bill stipulates), have been filed in Florida, Maryland, Ohio, Texas, Colorado, Massachusetts, Pennsylvania and New Jersey, although none have, as yet, been enacted. Other state bills appear to be imminent. Two federal bills were introduced in 1979--the "Truth in Testing Act of 1979," known as the Gibbons Bill or H.R. 3564, and the "Educational Testing Act of 1979," known as the Weiss Bill, or H.R. 4949. The former would cover achievement and occupational tests as well as admissions tests, but would not require total disclosure; the latter would be limited to admissions tests but would not require total disclosure.

All but two of the bills introduced apply to postsecondary education admissions testing only. They do not apply to standardized achievement tests used in public elementary and secondary schools, nor to personality, diagnostic, or minimal competency exams. An exception is the Massachusetts Bill which requires total disclosure of its competency tests. With the exception of the Gibbons Bill, these bills would not apply to occupational testing, civil service or licensing examinations. The New Jersey bill, however, would apply to all tests "developed by a test agency for the purpose of selection, placement, classification, graduation or any other bonafide reason concerning pupils in elementary and secondary, postsecondary or professional schools."

The arguments surrounding test disclosure legislation are compounded by disagreements about the role and power of testing companies and the quality of standardized tests used primarily for predicting student performance. Table 1 summarizes those arguments which deal with the issue of test disclosure.[2]

Issue 4: What is the meaning of test bias?

Perhaps the most difficult social, educational, technical, and legal issue facing educators in general and measurement specialists in particular, is the issue of test bias. Bias is such an important issue because it arises from our aspirations to achieve two highly valued goals. First, we have emerged from the 1970s with an ever growing awareness of the wide variety of cultures in our society and a desire to accommodate them. Second, we face the always present challenge of conducting good quality (fair and useful) assessment in our schools. These goals give rise to the need for testing methods that take into account cultural and linguistic differences in students.

Meeting both priorities is a difficult challenge because we often lack the combination of cultural or linguistic knowledge and test development skills required to do the

[2]The information in the table is taken from Searching for the Truth in "Truth in Testing" Legislation: A Background Report. Much of the above material has been abstracted from that report; those readers who wish to pursue the issues outlined are encouraged to obtain a copy of this publication. The report is available from ECS, 1860 Lincoln Street, Denver, Colorado 80295. The cost is $6.50 per copy.

TABLE 1

Debates For and Against Test Disclosure Legislation

Pro-Legislation Sentiments	Anti-Legislation Sentiments
Grade inflation, misuse have combined to give tests too much influence in admissions decisions.	Higher education's need for students has lessened importance of admissions' test scores.
A commitment to "truth in lending," "truth in advertising," sunshine laws and consumerism should extend to an area as important as admissions testing.	Test publishers and higher education institutions already provide ample information and protection; analogies to consumer movements are misleading.
Legislation will promote greater accuracy, validity of tests.	There are several competing public interests at stake; critics have not established an overriding need for legislation.
Legislation will encourage use of multiple criteria in selection process.	Legislation calling for full disclosure will lower the quality of tests.
The admissions test industry is not accountable to anyone.	Most institutions already use multiple criteria and test agencies encourage the practice.
Students can learn about testing and test strategy from examining test questions.	The industry is accountable to the psychometric profession, market forces, academic community.
Security need not be an issue; new measurement technology could enable testers to eliminate the problem.	Federal legislation would constitute dangerous, if not unconstitutional, federal incursion into education.
Development costs would not increase as much as testers suggest.	Legislation interferes with First Amendment right of colleges to determine who they want to teach.
Items now available only to expensive coaching schools would be available to everyone, benefiting poor students.	
There are many solutions to the comparability problem; the laws do not adversely affect comparability measurement.	
The fairness issue takes precedence over technical matters.	
Disclosure will help admissions officers as well as students.	

job. The equation is complex indeed. On one hand we have an examinee who brings to the test a language and set of cultural experiences that may represent any of hundreds of cultures. And, on the other hand, we have a test prepared by test makers (teachers or test publishers) who must make certain assumptions about language and cultural patterns in order to prepare test items. Claims are often made that tests are based on the language and culture of white, middle-class, suburban children and are inherently unfair to students who experience other cultural settings. Claims of ethnic, cultural, socio-economic and sex bias are widespread.

Currently, test publishers and educational researchers are devoting considerable effort to clarifying the definitions of and reasons for test bias, and to determine how to deal with its existence. For instance, in 1980 a National Symposium of Educational Research sponsored by Johns Hopkins University was devoted to the topic of test item bias methodology.

DEFINITIONS

Although no single technically correct definition of test bias exists, one which repeatedly appears in the writings of researchers and publishers is that a test is biased if individuals from different groups who are equally able, do not have equal probabilities of success. For example, on an achievement test, if students in one racial group score consistently lower than students from another group, and consistently lower than would be expected from their observed classroom performance, the test may be said to be biased against that group. Similarly, on a test used to select students for college admission, if students from one racial group score consistently lower than students from another group, but the performance of the two groups of students in the college program is comparable, the test may be said to be biased against the lower scoring group.

Several other definitions have been suggested. For example, one definition is that a test is biased if the different groups tested do not achieve the same average score on each item of the test. Another definition holds that a test is biased if two groups do not achieve similar total test scores. This definition allows for differences in performance on different items. These definitions assume that the groups are alike in knowledge of skills measured and any differences in performance are due to unfair items. These definitions have given rise to many public complaints of unfairness. However, it is critical to keep in mind that given our history of discriminatory educational practices, differences in performance may be caused by factors other than biased test items.

Another definition does not require that groups have the same ability or skill, but does require that differences hold true for all test items. That is, if differences are not uniform, it is assumed that the test items are measuring different things in the various groups.

Other kinds of bias are not inherent in the test but, rather, relate to how a test is used. For example, bias could be shown to occur if a test were used to make a selection decision simply because the test is correlated with a third variable that is relevant to and predictive of job performance even though the test itself has not been established as relevant to job performance. The use of a test could be biased if it assessed only one prerequisite skill and ignored equally predictive and important skills for which the pattern of group performance was noticeably different.

APPROACHES TO REDUCING TEST BIAS

It is important to point out that there is no clear-cut "solution" to the problem of test bias. No "culture-free" test has yet been devised, nor is the state of the art such that one can be developed. The best that can be done is for test-makers to make vigorous efforts to continuously screen tests for potential bias, and for test users to be sure that test results are used fairly in all cases.

One approach commonly used to avoid test bias is to have a panel of persons broadly representative of the various racial, ethnic and sexual groups that might be taking the test review the test questions. This helps ensure that test questions will not be biased or that they will not reflect only experiences or the culture of a particular group. This procedure should be undertaken not only when a test is first written, but periodically thereafter so that changes in our culture do not make some questions obsolete for some groups.

Another approach is to carefully examine the performance of various groups on the test as a whole as well as for individual questions. In this way, unusual variations in performance among the groups can be pinpointed, and the test questions reexamined in an effort to detect any characteristics or wording that would seem to make them biased towards a particular group. For publishers to conduct these studies, school districts must be willing to provide the demographic data necessary to perform the analyses.

Given the large number of languages and cultures in some educational environments, this process of careful test review and development will require significant time, money and patience.

Issue 5: What are the legal issues related to IQ testing?

People have and will probably continue to disagree about whether or how "intelligence" can be accurately and systematically measured. Some argue that evidence of intelligence can be reduced to a set of tasks which can be systematically measured through some form of performance or paper and pencil test. Others argue that traits such as common sense, wit, creativity, resourcefulness, ambition, and sensitivity are all important dimensions of intelligence and can never be adequately quantified in a test score.

IQ tests have historically been used to attempt to assess a child's aptitude for performance in school. These tests are designed to assess skills that are perceived to be prerequisites to learning skills such as verbal reasoning, spatial perception, etc. Thus, high scores on the tests are often used to place children in classes for the gifted. Conversely, low scores are often used to place children in special education classes for the mentally retarded. The most commonly used individually administered IQ tests, the Stanford-Binet and Wechsler Intelligence Scale for Children (WISC), are forms of "performance tests." Children are given a set of tasks to perform and are judged on the speed and accuracy with which they perform them. One important assumption behind the tests is that "intelligence" is distributed in society along a normal curve. This means that a small number of people in the society will be very bright or very dull, and the majority will cluster around a point defined as average intelligence.

Since the way in which IQ test scores are used has significant consequences for children (e.g., placement in classes for the retarded), legal challenges have focused both on the nature of the

tests and the ways in which the results are used. The most significant legal precedents in IQ testing come from a 1979 Federal District Court decision in a California case (Larry P. v. Riles, No. C71-2270 RFP, N.D. Cal. Decision 10/16/79) and a 1980 Federal District Court decision in an Illinois case (Parents in Action on Special Education v. Hannon, No. 74C3586, N.D. Ill. Decision 7/7/80).

The Larry P. v. Riles decision held that California school officials unlawfully discriminated against black children by using racially and culturally biased tests to classify and place them in classes for the educable mentally retarded (EMR). Judge Robert F. Peckham provides the following summary of his 131-page opinion.

This court finds in favor of plaintiffs, the class of black children who have been or in the future will be wrongly placed or maintained in special classes for the educable mentally retarded, on plaintiffs' statutory and state and federal constitutional claims. In violation of Title VI of the Civil Rights Act of 1964, the Rehabilitation Act of 1973, and the Education for All Handicapped Children Act of 1975, defendants have utilized standardized intelligence tests that are racially and culturally biased, have a discriminatory impact against black children and have not been validated for the purpose of essentially permanent placements of black children into educationally dead-end, isolated, and stigmatizing classes for the so-called educable mentally retarded. Further, these federal laws have been violated by defendants' general use of placement mechanisms that, taken together, have not been validated and result in a large over-representation of black children in the special E.M.R. classes.

"Defendants' conduct additionally has violated both state and federal constitutional guarantees of the equal protection of the laws. The unjustified toleration of disproportionate enrollments of black children in E.M.R. classes, and the use of placement mechanisms, particularly the I.Q. tests, that perpetuate those disproportions, provide a sufficient basis for the relief under the California Constitution. And under the federal Constitution, especially as interpreted by the Ninth Circuit Court of Appeals, it appears that the same result is dictated.

"Moreover, there is another basis for the federal constitutional ruling. Defendants' conduct, in connection with the history of I.Q. testing and special education in California, reveals an unlawful segregation intent. This intent was not necessarily to hurt black children, but it was manifested, inter alia, in the use of invalidated and racially and culturally biased placement criteria. This intent, consistent only with an impermissible and unsupportable assumption of higher incidence of mental retardation among blacks, cannot be allowed in the face of the constitutional prohibition of racial discrimination."

Relief granted to plaintiffs included an injunction against defendants' use of standardized intelligence tests for EMR identification or placement without court approval and an order that defendants monitor and eliminate disporportionate EMR placement of black children. The court decision also granted the reevaluation of all black children who were placed in EMR classes without the use of such tests, as well as supplemental education for all children found to have been misclassified.

The trigger for the Larry P. v. Riles court's legal scrutiny of IQ

tests and test bias was the disproportionate number of black children placed in EMR classes as a result of IQ tests and the serious injury of EMR placement to misclassified children. The court found that the EMR classes were "conceived of as 'dead-end classes'" for children incapable of learning the regular curriculum. Children in these classes tended to fall further and further behind children in regular classes since they were provided with instruction that deemphasized academic skills in favor of adjustment. Disproportionate numbers of black children had been placed in California's EMR classes. For example, the evidence showed that in the 20 districts accounting for 80 percent of the enrollment of black children in 1976-77, black students comprised about 27.5 percent of the student population and 62 percent of the EMR population. This disproportion cannot be explained by chance since "there is less than one in a million chance that the overenrollment of black children and the underenrollment of nonblack children in the EMR classes in 1967-77 would have resulted under a color-blind system of placement."

Although California law required IQ test scores to be "substantiated by" other evidence such as adaptive behavior (the ability to engage in social activities and perform everyday tasks), the court found that the "magic of numbers" was strong and that the available data suggested very strongly that the IQ scores were a pervasive influence in the placement process. The entire placement process often revolved around the demonstration of IQ.

In an introductory discussion of intelligence tests subtitled "The Impossibility of Measuring Intelligence," Judge Peckham noted that the expert testimony overwhelmingly rejected the concept that IQ was an objective measure of innate, fixed intelligence.

"Defendants' expert witnesses, even those closely affiliated with the companies that devise and distribute the standardized intelligence tests, agreed, with one exception, that we cannot truly define, much less measure, intelligence--I.Q. tests, like other ability tests, essentially measure achievement in skills covered by the examinations. The fact that IQ tests are developed according to the plausible but unproven assumption that intelligence is distributed in the population in accordance with a normal statistical curve--cautions us to look very carefully at what the tests do measure and exactly how they were validated for determining mental retardation."

Noting that the disparities in EMR placement of black children are also reflected historically in black performance in general on standardized intelligence tests, Judge Peckham examined three arguments used to explain the disparity in IQ scores--the genetic argument, the socio-economic argument, and cultural bias. Judge Peckham rejected the genetic argument because defendants were unwilling to admit any reliance on it for policy-making purposes and because the rather weak evidence in support of this explanation tends to rest on the disparities in the IQ scores, which overlooks possible bias in the tests themselves. Judge Peckham also rejected the socio-economic argument. Testimony and studies showed that the relatively low scores of black children do not result from mental disease attributable to the physical conditions of poverty. School performance, however, does vary somewhat according to socio-economic status.

On the other hand, Judge Peckham found the plaintiffs' evidence of racial and cultural bias in the IQ

tests more persuasive. "The first important inferential evidence is that the tests were never designed to eliminate cultural biases against black children; it was assumed, in effect, that black children were less 'intelligent' than whites." He later noted: "The tests had been adjusted, for example, to eliminate differences in the average scores between the sexes, but a comparable effort was not made and has never been made for black and white children."

The court also found that Wechsler's admission in 1944 (that the WISC's standardization was based upon white subjects only and that those norms cannot be used for the nonwhite population of the United States) applies with equal force to other standardized tests. These problems were not solved by the restandardization of the Stanford-Binet and WISC-R intelligence tests. The court went on to review a number of indicators that point to the existence of a cultural bias against black children's vocabulary and other linguistic differences, obviously biased items and more subtle kinds of bias involved in measuring knowledge of white culture. With only one exception, there was general agreement by all sides on the inevitable effect of cultural differences on IQ scores. Put succinctly by Professor Asa Hillard, black people have a "cultural heritage that represents an experience pool which is never used" or tested by the standardized IQ tests.

In analyzing the requirements of federal statutory law, the Larry P. v. Riles case set legal standards for validation of IQ tests used for EMR placement. Reviewing Title VI of the Civil Rights Act of 1964, the Rehabilitation Act of 1973, and the Education for All Handicapped Children Act of 1975 (EHA), and related case law, Judge Peckham concluded that the approach used in Title VII employment test cases was generally appropriate for allocating burden of proof for "validation" in the Larry P. v. Riles case. Under this procedure, tests shown to have a discriminatory impact cannot be utilized unless the employer is able to show that any given requirement has a manifest relationship to the employment in question. Judge Peckham noted, however, that the notion of predicting "job performance" cannot be effectively translated into an educational context given the differing purposes of employers and schools:

"Compulsory attendance of educational institutions is required by the state, and the schools are supposed to take children from different backgrounds and teach them the skills necessary for adaptation and success in our society. This points out a fundamental difference between the use of tests in employment and education, at least in the early years of schooling. If tests can predict that a person is going to be a poor employee, the employer can legitimately deny that person a job, but if tests suggest that a young child is probably going to be a poor student, the school cannot on that basis alone deny that child the opportunity to improve and develop the academic skills necessary to success in our society. Assignment to E.M.R. classes denies that opportunity through relegation to a markedly inferior, essentially dead-end, track."

Given this important distinction and federal regulations under EHA and the Rehabilitation Act requiring that tests and other evaluation materials be "validated for the specific purpose for which they are used," Judge Peckham replaced the predictive validity required in employment cases with an alternative kind of validation:

"We are not concerned now with predictions of performance, but rather whether the tests are validated with respect to the characteristics consistent with E.M.R. status and

placement in E.M.R. classes. E.M.R. classes exist 'for people whose mental capabilities make it impossible for them to profit from the regular educational program.' 'Mental retardation' is the touchstone, and retardation must make it 'impossible' to profit from the regular classes, even with remedial instruction. Defendents have the burden of showing validation of intelligence tests with respect to these characteristics."

In Parents in Action on Special Education v. Hannon, the presiding judge, Judge Grady, focused sharply on whether the IQ tests in question (WISC, WISC-R, and Stanford-Binet) are, in themselves, racially biased, and whether use of the tests as a part of the statute-mandated criteria for placement in classes of the "educable mentally handicapped" is racially discriminatory. In summary, the opinion concluded that:

(1) Only one item on the Stanford-Binet and a total of eight items on the WISC and WISC-R are culturally biased against black children, or at least sufficiently suspect that their use is inappropriate. These few items do not render the tests unfair and would not significantly affect the score of an individual taking the test.

(2) When used in conjunction with other statute-mandated criteria for determining an appropriate educational program for a child, these tests do not discriminate against black children in the Chicago schools.

In contrast to the Larry P. v. Riles decision, Judge Grady never reached the question of appropriate legal standards for evaluating compliance with federal law. Instead, Grady presented an exhaustive, item by item analysis of questions included in the three tests, found an insignificant number to be biased, and refused to enjoin Chicago's use of the tests as a part of the placement process.

The opinions in each of these cases are readable and informative. Readers interested in more detail and background on the opinions are encouraged to obtain and review copies of the opinions from the respective District Courts.

It is difficult to predict what will follow in the wake of these two opinions. While Judge Peckham in Larry P. v. Riles accepted the contention that IQ tests were biased, Judge Grady in Parents in Action v. Hannon rejected this allegation. Undoubtedly, further litigation will follow. The California Department of Education has already announced plans to appeal Larry P. v. Riles.

It is likely that the legal controversy over use of traditional IQ tests will spur research efforts to develop so-called "non-discriminatory" assessment batteries whose results will more accurately reflect the potential of minority children. One example of such a battery is the "System of Multicultural Pluralistic Assessment," known as SOMPA. SOMPA was developed by a sociologist at the University of California, Riverside, and is designed to provide a far broader picture of a child's potential based on a careful examination of the child's social and cultural background and experiences. It is unlikely that "alternative" IQ measures which are acceptable to critics of IQ tests will be developed and validated quickly.

Issue 6: What are the educational and legal issues surrounding minimum competency testing?

The fundamental purpose behind minimum competency testing is to determine whether students have acquired sufficient proficiency in certain basic and/or life skills to cope with the adult world. Two types of tests exist, tests that measure the basic academic skills of reading, writing and computation, and tests measuring "life skills" on topics such

as consumer awareness, health, citizenship, balancing a checkbook or applying for a bank loan.

In some states, the same test is given statewide, whereas in other states each district designs and administers its own test based on locally determined competence.

A 1979 study sponsored by the National Institute of Education investigated 31 state and 20 local district competency testing programs in the United States. An executive summary of that study states:

"Sixteen of the 31 state-level programs were mandated by the State Board of Education, and 15 were initiated by the state legislature. Two of the legislated mandates call for temporary programs; one State Board-initiated program and one legislated program permit voluntary participation of local school districts. Two other states emphasize the competency-based instructional aspects of their programs rather than the testing components.

"Of the 20 local programs studied, five developed in states without statewide requirements for minimum competency testing. Of the remaining 15 districts, eight began instituting minimum competency testing programs prior to state mandates, while seven districts implemented programs in response to such mandates.

"The majority of programs, both state and local, were developed in the two to three years since 1976, but the age of programs ranged from 18 years to less than one year with ongoing pilot-testing. Fourteen state programs have been fully implemented, while 17 are being phased in. For example, many state programs are introducing new graduation requirements or curriculum changes over a period of years and hence, these programs will not be "in place" until some time in the future. By comparison, 13 of the 20 local programs have already been fully implemented, while seven programs are phasing in mandated changes.

"Programs in only four states have had litigation associated with them in any way--Delaware, Florida, Maryland, and North Carolina--and the majority of this activity has occurred in Florida.

"With respect to goals and purposes, 14 states cited certification of basic skills competency prior to high school graduation as a major purpose, and two states reported using competency achievement as one criterion for grade-to-grade promotion as a reason for implementing a minimum competency testing program. The most frequently cited purpose for instituting such a program was to identify students in need of remediation; 19 states reported this purpose. Curriculum improvement was mentioned by 10 states as a major program goal. By comparison, 16 local districts reported certification of basic skills as one reason for developing a minimum competency testing program; four districts cited the use of test results, along with other information, to determine grade-to-grade promotion as a major purpose of the program. Eleven programs reported purposes related to providing remediation and seven districts mention curriculum change as a major purpose behind program implementation.

"Reading and mathematics were competency areas assessed in all state and local programs. Twenty-seven of the state programs assessed skills in language arts and/or writing, while 15 local districts assess these same skills. Skills in other subject areas, such as speaking, listening, consumer economics, science, government, and history, are assessed in only a few programs. Almost all of the tests administered in both state and local programs consist primarily of multiple-choice items, and a writing sample is the most frequently

selected non-multiple-choice assessment."[3]

LEGAL ISSUES

Many of the legal issues involved in competency testing are inextricably linked to issues of test quality and the quality of educational programs designed to support competency testing. For example, the nature and quality of a competency test may trigger legal challenge, but test quality is and should be in itself an educational issue. Similarly, insuring quality and effectiveness in basic and remedial instructional programs is one of the central missions of education. Nevertheless, in examining minimum competency programs, courts are likely to closely examine these instructional activities. While it seems impossible to clearly disentangle "legal" from "educational" issues in minimum competency testing, it is useful to review the issues courts have examined to date.

The distinction between using a competency test only as a diagnostic tool to identify student weaknesses in basic skills and tying high school graduation to successful performance on the test, is crucial in examining the legal implications of minimum competency testing. The legality of a testing program will usually depend more on how the test results are used than on the nature of the test itself. For example, as McClung points out in a legal review of competency testing:

"Using the test results as the primary basis for any decision that will cause serious harm to a student raises the initial legal questions. The trigger for legal analysis is this injury. Assuming there is injury, the following questions arise: Who is responsible for that injury and does that person or agency have sufficient justification for causing that injury?

"If there is no injury, then there is no legal problem. Competency tests can be used in many ways that cause no injury to a student. For example, competency tests could be used simply to determine the general level of student performance in basic skills on a statewide or district level; to identify basic skill areas in an instruction program that need more emphasis; or to diagnose areas in which an individual student needs specific help. In such cases, there is usually no injury and no legal problem.

"On the other hand, competency tests can be used to make decisions about individual students that have potential for grave injury. For example, competency tests can be used for tracking, grade promotion, or denial of a regular high school diploma. Diploma denial, as mandated in Florida and California, probably causes the greatest injury to an individual student, and therefore raises the most serious legal questions (p. 657-658)."[4]

Minimum competency testing requirements that incorporate some sanction upon students for failing to pass the tests run the greatest risk of legal challenge. These legal challenges are most likely to be raised if competency testing programs touch on any of the following issues:

---

[3]Gorth, W.P., and Perkins, M.R., A Study of Minimum Competency Testing Programs: Final Summary and Analysis Report. Amherst, MA: National Evaluation Systems, Inc., December 1979.

[4]McClung, M.S., "Competency testing programs: Legal and educational issues," Fordham Law Review, 47, 1979, 651-711.

- Potential for racial and linguistic discrimination
- Adequacy of advance notice and phase-in periods prior to the initial use of the test as a graduation requirement
- Psychometric validity or reliability of the tests
- Match between the instructional program and the test
- The degree to which remedial instruction may create or reinforce tracking

1. Potential for racial and linguistic discrimination. Briefly stated, some states and many local school districts in the past have been found to have discriminated against racial and linguistic minority students in violation of the equal protection clause of the U.S. Constitution and Title VI of the Civil Rights Act of 1964. Examples of such states and districts include those that have been held by courts to have operated "dual school systems" for blacks and whites and who have been ordered to desegregate, and those that have been found not to be providing adequate bilingual instruction in accord with the U.S. Supreme Court's ruling Lau v. Nichols. In states or districts which have been subject to or are vulnerable to such findings, the effect of minimum competency testing requirements may be to reinforce the effects of prior discrimination. That is, the minimum competency testing sanction could pile one injury (diploma denial) on top of another (prior denial of equal educational opportunity).

2. Adequacy of advance notice and phase-in periods prior to the initial use of the test as a graduation requirement. Legal concerns for fairness and due process will require extensive notice of minimum competency testing requirements to students and parents. For example, the first class of students subject to a minimum competency testing requirement might not know that passing a competency test will be a condition for acquiring a diploma. The school district, in fact, would have explicitly approved students' progress by promoting them each year even though many of them lacked basic skill proficiencies. It is also likely that many, if not most, of those students failing the test might have studied differently and teachers taught differently had they received advance notice of the requirement.

Procedures for notifying students vary from school to school. In most districts students are first given general notice of the proficiency requirement for a diploma and then at a later date notified of the specific performance objectives to be measured by the proficiency test. Students, parents and teachers should be given notice of both performance objectives and assessment procedures as soon after their adoption as possible.

Traditional notions of due process require adequate prior notice of any rule that could cause irreparable harm to a person's educational or occupational prospects. Notification of requirements after completing most of one's educational program may be viewed as both unfair and inadequate, especially if the minimum competency test is designed to measure knowledge and skills not previously taught in the district's classrooms.

3. Psychometric validity or reliability of minimum competency tests. All tests ought to meet reasonable professional psychometric standards of validity and reliability. Simply stated, validity refers to whether or not a test measures what it purports to measure, and reliability refers to whether or not the test measures student performance accurately from one test administration to another. The most widely

accepted professional test development standards are the Standards for Educational and Psychological Tests, published by the American Psychological Association. It is likely that minimum competency tests will be subjected to careful scrutiny against such benchmarks as the Standards.

4. Match between the instructional program and the test. Most persons would agree that fairness requires that a school's curriculum and instruction be matched to the competencies measured by a test. In other words, the test would be unfair if it attempted to measure what the school did not teach. This concept should be considered in terms of both curriculur validity and instructional validity.

Curricular validity is a measure of how well test items match the objectives of the curriculum. An analysis of curricular validity would require comparison of the test objectives with the school's stated course objectives. This becomes important, for example, if the curriculum is not specifically designed to teach functional competency and the use of a test covering functional competency is considered. It might be unfair to deny students their diplomas because they did not learn these functional competencies. In such a situation, failure on the minimum competency test might indicate that the school did not offer an appropriate curriculum.

A minimum competency test should also have what may be called instructional validity. Even if the curricular objectives of the school correspond to those of the competency test, there might be a discrepancy between the stated objectives of the school and what is actually being taught in the classroom. Instructional validity obviously does not require prior exposure of the student to the exact questions asked on the test, but it does require exposure to the kind of knowledge and skills that would enable a student to answer the test questions.

It is important to note that content validity does not ensure either curricular or instructional validity. They are related, but distinguishable concepts. Content validity is a measure of how well test items represent the body of skills and knowledge that the test purports to measure but is not necessarily a measure of how well the test items represent either a school's curricular objectives or instruction. Instructional validity should be the central concern because content and curricular validity mean very little if the test items are not representative of instruction actually received by the student.

5. The degree to which remedial instruction may create or reinforce tracking. Most minimum competency testing programs implicitly or explicitly require remedial instruction for students found to be deficient in basic skills. In districts subject to findings of prior racial or linguistic discrimination as described above, one effect of minimum competency testing requirements may be to inappropriately channel or "track" disproportionate numbers of minority students into remedial programs on the basis of their test results. This could have the effect of "resegregating" students into remedial programs in direct contradiction to prior orders to desegregate school systems.

THE DEBRA P. v. TURLINGTON DECISION

To date, the only major legal challenge to competency testing was mounted in Florida. In Debra P. v. Turlington, a group of black student plaintiffs sued the state in Federal Distict Court to have the state's competency testing program ruled unconstitutional. Plaintiffs challenged the test on each of the grounds mentioned above.

In July 1979, the court held that Florida's competency testing program did not give all students adequate notice of the inclusion of the competency test as a graduation requirement, and that the competency testing program carried forward the effects of prior racial discrimination in violation of the due process and equal protection clauses of the Fourteenth Amendment of the U.S. Constitution, Title VI of the Civil Rights Act of 1964, and the Equal Educational Opportunities Act of 1974. As a remedy, the court enjoined Florida from using the test as a diploma requirement for four years, until the 1982-83 school year. The court did not, however, deny use of the competency test during this four-year period for assessing the effects of instruction.

Although the court found psychometric deficiencies in Florida's test, it did not find these deficiencies to be unconstitutional. The court did not address in any depth the issue of the correlation between the test and instructional program.

Issue 7: Are tests being used to evaluate teachers in schools?

Tests are being used to evaluate teachers in a variety of ways. But tests are never used as the sole criterion of teacher evaluation because of the complexity of the learning process. Since many factors influence learning, some under teacher control and some not, teacher evaluation must be done very carefully.

The types of test scores that can play a role in teacher evaluation are the achievement test scores of students, test scores of licensing examinations, and the scores of tests used in the teacher selection and hiring processes.

The evaluation of teachers by using the achievement test scores of the students they teach is a very delicate process. If a group of students who have previously shown patterns of growth in test scores do not grow over an extended period of time, and this phenomenon is apparent in the test scores of all or nearly all students in the group with the same teacher, then those test scores can be combined with other information about the teacher as part of the teacher evaluation process. However, if test scores of students are to be used in this way, they must be used very carefully and with full awareness of the potential difficulties with this evaluation strategy.

The first difficulty is that factors apart from the school experience can greatly influence student achievement. Since teachers have no control over many of these factors they cannot be held accountable. For example, characteristics such as the child's ability to learn, and the child's motivation, are not totally within the teacher's control. The student's home environment also exerts great influence on learning. In fact, some research suggests that some non-school factors may far outweigh school factors in determining achievement. When these factors begin to interact with the various characteristics of the school learning environment, it becomes difficult to sort out the component of learning that is influenced by the teacher and the components that are influenced by non-school conditions.

The second difficulty with using student test scores to evaluate teacher performance is the complexity of the desired end product. In school, teachers endeavor to help the child to gain knowledge and skills in many academic areas, some common to all students, some unique to an individual student. In addition, teachers attempt to develop values, attitudes and interpersonal skills that will benefit a student in society. Given all of these desired

traits along with the complexity and uniqueness of each individual student, it becomes impossible to define the characteristics of the "desired" end product to evaluate.

Even when it is possible to define the citizen we want our schools to produce, we have great difficulty reflecting many of the important characteristics in reliable and valid test scores. Though we can use tests to document some of the basic achievement areas, the focus of these tests is very broad and general and may not reflect the important educational objectives in a given school district, building, or classroom. Furthermore, other desired outcomes, such as attitudes, values and interpersonal skills are inherently complex and not easily measured in an objective way in school settings or otherwise.

The third potential difficulty with using student test scores for teacher evaluation is that learning does not take place at a steady and predictable rate. Even if we could define and measure the end product of schooling and control most of the factors that influence that product, we could not assume that every child would gain new knowledge and skills at the same pace. Some would learn faster than others. Some would grow slowly then spurt ahead--all according to the nature of human development. This fact must be taken into account in evaluating teacher performance via student test scores.

State licensing examinations are also used as a form of teacher evaluation. Though most states issue licenses on the basis of the completion of specified college courses or degrees, some also include an examination as part of the credentialling process.

In the field of education, tests have been in use for decades for certifying teacher competence. The State of South Carolina, for example, has used the National Teacher Examinations (NTE) to certify teachers since 1945. The Education Commission of the States[5] has developed an excellent summary of the current status of such testing.

The National Teacher Examinations, which are published and administered by the Educational Testing Service, include examinations covering academic preparation in professional education and general education (writing, science, math, social studies, literature) as well as academic preparation in 26 subject-field specializations. The tests typically focus on the recall of factual information with some use of higher order mental operations tests as well.

In the fall of 1977, four states required or recommended use of NTE results for initial certification purposes. These states were Mississippi, North Carolina, South Carolina and West Virginia. Louisiana was added to this list in 1978. In addition to these five states, at least 23 states used the NTE for special purposes, ranging from obtaining statewide data for teacher education studies (Alabama) to validating credits earned at nonaccredited institutions (California, Delaware). In June 1978, the Florida Legislature passed a bill requiring, in part, a test of teaching competency and subject matter mastery for initial certification. Working steadily over a period of four to five years, the Georgia State Department of Education developed test instruments for a "Performance Based Teacher Certification" program, and first administered the test in November 1978. In early 1979, hearings were held in North Carolina on plans for a "Quality Assurance Program for Professional Personnel" in which testing for

---

[5]Vlaanderen, R. "Trends in competency-based teacher certification." Denver, CO: Education Commission of the States, March 1980.

teaching competencies and subject matter mastery plays a major role in the certification process. The program was adopted in the fall of 1979.

In 1979, several state legislatures introduced bills embodying the testing concept in teacher certification. In Arkansas, a bill was passed in record time, while similar bills in Colorado, Kansas, Arizona, Missouri and Vermont died in committee. Bills were introduced in Alabama, Iowa and Oklahoma in 1980 and again, in a special session, in Arizona. State Board action has mandated testing in Alabama and Tennessee.

Test scores are also used, in some instances, when several teachers are being considered for a limited number of teaching positions. The employers may use a test as part of the selection process. In this case, all teachers may be certified, but another test might be used to determine knowledge of subject matter and/or ability to perform in a certain educational environment. As in the other instances, test scores should never be the only criteria considered in the selection process. But they can be a valuable selection aid when used carefully with other performance information.

Annotated Bibliography

Test Purposes and Users

Anderson, B.L., Stiggins, R.J., and Hiscox, S.B. Guidelines for selecting basic skills and life skills tests. Portland, OR: Northwest Regional Educational Laboratory, 1980.

This short guide designed for teachers and administrators discusses test purposes and characteristics to consider when selecting tests. Lists of currently available basic skills and life skills tests are provided along with the names and addresses of test publishers.

Brown, F.G. Guidelines for test use: A commentary on the standard for educational and psychological tests. Washington, D.C.: National Council on Measurement in Education, 1980

This book is designed for teachers, counselors, school psychologists, administrators, parents and others concerned with educational measurement. It is a nontechnical explanation of the Standards.

Burrill, L.E. How a standardized achievement test is built, test service notebook 125. New York, NY: The Psychological Corporation.

The steps described are typical of the way tests are built by many major test publishers. Other short articles on related topics are available from The Psychological Corporation, New York, NY 10017.

Feder, B. The complete guide to taking tests. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1979.

This book is written for test takers who want to take some of the mystery out of testing.

Parents and testing. Washington, D.C.: National Education Association, 1979.

This guide provides parents with information on how they should and can be involved with schools' testing programs. It also gives the NEA position on student testing.

Rebell, M.A., and Block, A.R. Competence assessment and the courts: An overview of the state of the law. Boston, MA: McBer, 1980.

This study looks at the implication of legal cases on a wide variety of educational testing situations, including certification, IQ tests, ability tracking, and graduate school admissions tests.

Teachers and testing. Washington, D.C.: National Education Association, 1979.

Teachers are provided with general informaton on how and why tests are used as well as their strengths and weaknesses. The NEA resolutions relating to testing issues are given.

Achievement Test Score Decline

Munday, L. Declining admissions test scores. ACT Research, Report #71, Iowa City, IA: The American College Testing Program, 1976.

Several indices of declining academic achievement are summarized. However, the principle focus is on declining ACT Assessment Program test scores. Correlates of score decline are identified and potential explanations are explored.

Harnischfeger, A., and Wiley, D.E. Achievement test score decline: Do we need to worry? Monograph of CEMREL, Inc., 3120 59th Street, St. Louis, MI 63139, 1976.

This 160-page monograph reviews several potential explanations for declining academic achievement test scores. Data are presented in association with the potential explanations presented and conclusions are drawn regarding each explanation. An excellent summary of conclusions is presented.

College Entrance Examination Board. On further examination. New York, NY: 1977.

This monograph reports the results of the deliberations of the CEEB Advisory Panel on the Scholastic Aptitude Test Score Decline. Potential explanations related to school and nonschool factors are examined and accepted or rejected as viable. Conclusions are presented regarding multiple causes.

Truth in Testing

Brown, R. Searching for the truth in "truth in testing" legislation: A background report. Denver, CO: Education Commission of the States, 1980.

This is a readable summary of the background and current issues in truth in testing. It also summarizes relevant pending federal and state legislation.

Brown, R. Searching for the truth about truth in testing. Compact, Winter, 1980, 7-11.

This article is a much abbreviated summary of the issues presented in the background report listed above.

Nairn, A. and Associates. The reign of ETS: The corporation makes up minds. Washington, D.C., 1980.

The Nairn report on ETS was sponsored by Ralph Nader and offers a strong indictment of many of ETS' practices.

Educational Testing Service. Test scores and family income. Princeton, NJ, February 1980.

Educational Testing Service. Test use and validity. Princeton, NJ, February 1980.

The two ETS reports were developed in response to the Nairn report.

Cultural Bias

Burrill, L.E. and Wilson, R. Fairness and the matter of bias: Test service notebook 36. New York, NY: The Psychological Corporation, 1980.

This article succintly covers major issues in facial bias, item bias and bias in selection and prediction.

Burrill, L.E. Statistical evidence of potential bias in items and tests assessing current educational status. Paper presented at the Fourteenth Annual Southeastern Conference on Measurement in Education, 1975.

This paper describes various definitions and interpretations of bias and provides a useful reference list.

Sheppard, L., Camilli, G., and Averill, M. Comparison of six procedures for detecting test item bias using internal and external ability criteria. A paper presented to the National Council on Measurement in Education Annual Meeting, Boston, 1980.

This paper not only provides a thorough comparison of procedures for detecting test item bias, but also contains an extensive reference list to the literature on test item bias.

IQ Testing

Larry P. v. Riles, No. C71-2270 RFP, N.D. Cal. Decision 10/16/79.

Readers who are interested in pursuing the issues raised in the Larry P. decision are urged to obtain a transcript of the decision and read it in its entirety. The decision is readable, to the point, and appropriate for a lay reader.

Parents in Action on Special Education vs. Hannon. No. 74C3586, N.D. Ill., Decision 7/7/80.

This transcript of the Parents in Action case provides a detailed item by item analysis of the IQ tests in question.

Notes on Larry P. Footnotes (Newsletter of the Law and Education Center, Education Commission of the States, Denver, CO) Vol. 1, No. 2, Spring 1980.

This newsletter presents a short, readable analysis of the Larry P. v. Riles case.

Minimum Competency Testing

Bunda, M.A., and Sanders, J.R. (Eds.) Practices and problems in competency based measurement. Washington, D.C.: National Council on Measurement in Education, 1979.

This 144 page book provides articles on the key issues in competency based testing.

Debra P. v. Turlington. Footnotes (Newsletter of the Law and Education Center, Education Commission of the States, Denver, CO) Vol. 1, No. 1, November 1979.

This newsletter provides a short readable review of the key issues in the Debra P. v. Turlington case.

Gorth, W.P. and Perkins, M.R. A study of minimum competency testing programs: Final summary and analysis report. Amherst, MA: National Evaluation Systems, 1979.

This report summarizes the current status of the implementation of minimum competency testing across the country.

McClung, M.S. Competency testing programs: Legal and educational issues. Fordham Law Review, 1979, 47, 651-711.

This article is an exhaustive review of legal issues which incorporates potential implications of the Debra P. vs. Turlington decision.

Shoemaker, J.S. Minimum competency testing: Implications for instruction. Washington, D.C.: National Institute of Education, January 1979.

This paper presents a discussion of design considerations in the development of minimum competency testing programs that will maximize the utility of the program for instructional uses.

Rosewater, A. Minimum competency testing programs and handicapped students: Perspectives on policy and practice. Washington, D.C.: George Washington University Institute for Educational Leadership, 1979.

This paper presents a review of policy and practical problems involved in implementing minimum competency testing programs for the handicapped.

Teacher Testing and Evaluation

The Psychological Corporation. Summaries of court decisions on employment testing, 1968-1977. New York, NY, 1978.

This book summarizes court decisions on employment testing in both the private and public sector. It is not limited to educational personnel.

Vlaanderen, R. Trends in competency based teacher certification. Denver, CO: Education Commission of the States, March 1980.

This paper presents a summary of the current status of teacher competency testing.

Appendix A

A Glossary of Measurement Terms

The following glossary is used with the permission of the Psychological Corporation, New York, N.Y. 10017

Similar glossaries may be obtained from other major test publishers.

Test Service Notebook 13

A Glossary of Measurement Terms

BLYTHE C. MITCHELL, *Consultant, Test Department*

This glossary of terms used in educational and psychological measurement is primarily for persons with limited training in measurement, rather than for the specialist. The terms defined are the more common or basic ones such as occur in test manuals and educational journals. In the definitions, certain technicalities and niceties of usage have been sacrificed for the sake of brevity and, it is hoped, clarity.

The definitions are based on the usage of the various terms as given in the current textbooks in educational and psychological measurement and statistics, and in certain specialized dictionaries. Where there is not complete uniformity among writers in the measurement field with respect to the meaning of a term, either these variations are noted or the definition offered is the one that the writer judges to represent the "best" usage.

academic aptitude. The combination of native and acquired abilities that are needed for school learning; likelihood of success in mastering academic work, as estimated from measures of the necessary abilities. (Also called *scholastic aptitude, school learning ability, academic potential*)

achievement test. A test that measures the extent to which a person has "achieved" something, acquired certain information, or mastered certain skills – usually as a result of planned instruction or training.

age norms. Originally, values representing typical or average performance for persons of various *age* groups; most current usage refers to sets of complete score interpretive data for appropriate successive age groups. Such norms are generally used in the interpretation of mental ability test scores.

alternate-form reliability. The closeness of correspondence, or correlation, between results on alternate (i.e., equivalent or parallel) forms of a test; thus, a measure of the extent to which the two forms are consistent or reliable in measuring whatever they do measure. The time interval between the two testings must be relatively short so that the examinees themselves are unchanged in the ability being measured. See RELIABILITY, RELIABILITY COEFFICIENT.

anecdotal record. A written description of an incident in an individual's behavior that is reported objectively and is considered significant for the understanding of the individual.

aptitude. A combination of abilities and other characteristics, whether native or acquired, that are indicative of an individual's ability to learn or to develop proficiency in some particular area if appropriate education or training is provided. Aptitude tests include those of general academic ability (commonly called mental ability or intelligence tests); those of special abilities, such as verbal, numerical, mechanical, or musical; tests assessing "readiness" for learning; and prognostic tests, which measure both ability and previous learning, and are used to predict future performance – usually in a specific field, such as foreign language, shorthand, or nursing.

Some would define "aptitude" in a more comprehensive sense. Thus, "musical aptitude" would refer to the combination not only of physical and mental characteristics but also of motivational factors, interest, and conceivably other characteristics, which are conducive to acquiring proficiency in the musical field.

arithmetic mean. A kind of average usually referred to as the *mean*. It is obtained by dividing the sum of a set of scores by their number.

average. A general term applied to the various measures of central tendency. The three most widely used averages are the arithmetic mean (mean), the median, and the mode. When the term "average" is used without designation as to type, the most likely assumption is that it is the *arithmetic mean*.

battery. A group of several tests standardized on the same sample population so that results on the several tests are comparable. (Sometimes loosely applied to any group of tests administered together, even though not standardized on the same subjects.) The most common test batteries are those of school achievement, which include subtests in the separate learning areas.

bivariate chart (bivariate distribution). A diagram in which a tally mark is made to show the scores of one individual on *two variables*. The intersection of lines determined by the horizontal and vertical scales form cells in which the tallies are placed. Such a plot provides frequencies for the two distributions, and portrays the relation between the two variables as a basis for computation of the product-moment correlation coefficient.

ceiling. The upper limit of ability that can be measured by a test. When an individual makes a score which is at or near the highest possible score, it is said that the test has too low a "ceiling" for him; he should be given a higher level of the test.

central tendency. A measure of central tendency provides a single most typical score as representative of a group of scores; the "trend" of a group of measures as indicated by some type of average, usually the *mean* or the *median*.

coefficient of correlation. A measure of the degree of relationship or "going-togetherness" between two sets of measures for the same group of individuals. The correlation coefficient most frequently used in test development and educational research is that known as the Pearson or *product-moment r*. Unless otherwise specified, "correlation" usually refers to this coefficient, but *rank*, *biserial*, *tetrachoric*, and other methods are used in special situations. Correlation coefficients range from .00, denoting a complete absence of relationship, to +1.00, and to −1.00, indicating perfect positive or perfect negative correspondence, respectively. See CORRELATION.

composite score. A score which combines several scores, usually by addition; often different weights are applied to the contributing scores to increase or decrease their importance in the composite. Most commonly, such scores are used for *predictive* purposes and the several weights are derived through multiple regression procedures.

concurrent validity. See VALIDITY (2).

construct validity. See VALIDITY (3).

content validity. See VALIDITY (1).

correction for guessing (correction for chance). A reduction in score for wrong answers, sometimes applied in scoring true-false or multiple-choice questions. Such scoring formulas (R − W for tests with 2-option response, R − ½W for 3 options, R − ⅓W for 4, etc.) are intended to discourage guessing and to yield more accurate rankings of examinees in terms of their true knowledge. They are used much less today than in the early days of testing.

correlation. Relationship or "going-togetherness" between two sets of scores or measures; tendency of one score to vary concomitantly with the other, as the tendency of students of high IQ to be above average in reading ability. The existence of a strong relationship — i.e., a high correlation — between two variables does not necessarily indicate that one has any causal influence on the other. See COEFFICIENT OF CORRELATION.

criterion. A standard by which a test may be judged or evaluated; a set of scores, ratings, etc., that a test is designed to measure, to predict, or to correlate with. See VALIDITY.

criterion-referenced (content-referenced) test. Terms often used to describe tests designed to provide information on the specific knowledge or skills possessed by a student. Such tests usually cover relatively small units of content and are closely related to instruction. Their scores have meaning in terms of *what* the student knows or can do, rather than in their relation to the scores made by some external reference group.

criterion-related validity. See VALIDITY (2).

culture-fair test. So-called culture-fair tests attempt to provide an equal opportunity for success by persons of all cultures and life experiences. Their content must therefore be limited to that which is equally common to all cultures, or to material that is entirely unfamiliar and novel for all persons whatever their cultural background. See CULTURE-FREE TEST.

culture-free test. A test that is free of the impact of all cultural experiences; therefore, a measure reflecting only hereditary abilities. Since culture permeates all of man's environmental contacts, the construction of such a test would seem to be an impossibility. Cultural "bias" is not eliminated by the use of non-language or so-called performance tests, although it may be reduced in some instances. In terms of most of the purposes for which tests are used, the validity (value) of a "culture-free" test is questioned; a test designed to be equally applicable to all cultures may be of little or no practical value in any.

curricular validity. See VALIDITY (2).

decile. Any one of the nine points (scores) that divide a distribution into ten parts, each containing one-tenth of all the scores or cases; every tenth percentile. The first decile is the 10th percentile, the eighth decile the 80th percentile, etc.

deviation. The amount by which a score differs from some reference value, such as the mean, the norm, or the score on some other test.

deviation IQ (DIQ). An age-based index of general mental ability. It is based upon the difference or deviation between a person's score and the typical or average score for persons of his chronological age. Deviation IQs from most current scholastic aptitude measures are standard scores with a mean of 100 and a standard deviation of 16 for each defined age group.

diagnostic test. A test used to "diagnose" or analyze; that is, to locate an individual's specific areas of weakness or strength, to determine the nature of his weaknesses or deficiencies, and, wherever possible, to suggest their cause. Such a test yields measures of the components or subparts of some larger body of information or skill. Diagnostic achievement tests are most commonly prepared for the skill subjects.

difficulty value. An index which indicates the percent of some specified group, such as students of a given age or grade, who answer a test item correctly.

discriminating power. The ability of a test item to differentiate between persons possessing much or little of some trait.

discrimination index. An index which indicates the *discriminating power* of a test item. The most commonly used index is derived from the number passing the item in the highest 27 percent of the group (on total score) and the number passing in the lowest 27 percent.

distractor. Any incorrect choice (option) in a test item.

distribution (frequency distribution). A tabulation of the scores (or other attributes) of a group of individuals to show the number (frequency) of each score, or of those within the range of each interval.

equivalent form. Any of two or more forms of a test that are closely parallel with respect to the nature of the content and the number and difficulty of the items included, and that will yield very similar average scores and measures of variability for a given group. (Also referred to as *alternate, comparable,* or *parallel* form.)

error of measurement. See STANDARD ERROR OF MEASUREMENT.

expectancy table ("expected" achievement). A term with two common usages, related but with some difference:

(1) A table or other device for showing the relation between scores on a predictive test and some related outcome. The outcome, or criterion status, for individuals at each level of predictive score may be expressed as (a) an average on the outcome variable, (b) the percent of cases at successive levels, or (c) the probability of reaching given performance levels. Such tables are commonly used in making predictions of educational or job success.

(2) A table or chart providing for an interpretation of a student's obtained score on an achievement test with the score which would be "expected" for those at his grade level and with his level of scholastic aptitude. Such "expectancies" are based upon actual data from administration of the specified achievement and scholastic aptitude tests to the same student population. The term "anticipated" is also used to denote achievement as differentiated by level of "intellectual status."

extrapolation. In general, any process of estimating values of a variable beyond the range of available data. As applied to test norms, the process of extending a norm line into grade or age levels not tested in the standardization program, in order to permit interpretation of extreme scores. Since this extension is usually done graphically, considerable judgment is involved. Extrapolated values are thus to some extent arbitrary; for this and other reasons, they have limited meaning.

f. A symbol denoting the *frequency* of a given score or of the scores within an interval grouping.

face validity. See VALIDITY (1).

factor. In mental measurement, a hypothetical trait, ability, or component of ability that underlies and influences performance on two or more tests and hence causes scores on the tests to be correlated. The term "factor" strictly refers to a theoretical variable, derived by a process of *factor analysis* from a table of intercorrelations among tests. However, it is also used to denote the psychological interpretation given to the variable—i.e., the mental trait assumed to be represented by the variable, as verbal ability, numerical ability, etc.

factor analysis. Any of several methods of analyzing the intercorrelations among a set of variables such as test scores. Factor analysis attempts to account for the interrelationships in terms of some underlying "factors," preferably fewer in number than the original variables, and it reveals how much of the variation in each of the original measures arises from, or is associated with, each of the hypothetical factors. Factor analysis has contributed to an understanding of the organization or components of intelligence, aptitudes, and personality; and it has pointed the way to the development of "purer" tests of the several components.

forced-choice item. Broadly, any multiple-choice item in which the examinee is *required* to select one or more of the given choices. The term is most often used to denote a special type of multiple-choice item employed in personality tests in which the options are (1) of equal "preference value," i.e., chosen equally often by a typical group, and are (2) such that one of the options discriminates between persons high and low on the factor that this option measures, while the other options measure other factors. Thus, in the *Gordon Personal Profile,* each of four options represents one of the four personality traits measured by the *Profile,* and the examinee must select both the option which describes him *most* and the one which describes him *least.*

frequency distribution. See DISTRIBUTION.

g. Denotes *general* intellectual ability; one dimensional measure of "mind," as described by the British psychologist Spearman. A test of "*g*" serves as a general-purpose test of mental ability.

grade equivalent (GE). The grade level for which a given score is the real or estimated average. Grade-equivalent interpretation, most appropriate for elementary level achievement tests, expresses obtained scores in terms of *grade* and *month of grade,* assuming a 10-month school year (e.g., 5.7). Since such tests are usually standardized at only one (or two) point(s) within each grade, grade equivalents between points for which there are data-based scores must be "estimated" by *interpolation.* See EXTRAPOLATION, INTERPOLATION.

grade norms. Norms based upon the performance of pupils of given grade placement. See GRADE EQUIVALENT, NORMS, PERCENTILE RANK, STANINE.

group test. A test that may be administered to a number of individuals at the same time by one examiner.

individual test. A test that can be administered to only one person at a time, because of the nature of the test and/or the maturity level of the examinees.

intelligence quotient (IQ). Originally, an index of brightness expressed as the ratio of a person's mental age to his chronological age, MA/CA, multiplied by 100 to eliminate the decimal. (More precisely – and particularly for adult ages, at which mental growth is assumed to have ceased – the ratio of mental age to the mental age normal for chronological age.) This quotient IQ has been gradually replaced by the deviation IQ concept.

It is sometimes desired to give additional meaning to IQs by the use of verbal descriptions for the ranges in which they fall. Since the IQ scale is a continuous one, there can be no inflexible line of demarcation between such successive category labels as very superior, superior, above average, average, below average, etc.; any verbal classification system is therefore an arbitrary one. There appears to be, however, rather common use of the term *average* or *normal* to describe IQs from 90-109 inclusive.

An IQ is more definitely "interpreted" by noting the normal percent of IQs within a range which includes the IQ, and/or

[**intelligence quotient (IQ)**, continued.]

by indicating its percentile rank or stanine in the total national norming sample. Column 2 of Table 1 shows the normal distribution of IQs for M = 100 and S.D. = 16, showing percentages within successive 10-point intervals. (For IQs whose S.D. is greater than 16, the percentages for the extreme IQ ranges will be larger, and those for IQs near the mean will be smaller, than those shown in the table.) Table 1 indicates that 47 percent, approximately one-half of "all" persons, have IQs in the 20-point range of 90 through 109; an IQ of 140 or above would be considered as extremely high, since fewer than one percent (0.6) of the total population reach this level, and fewer than one percent have IQs below 60. From the cumulative percents given in Column 3, it is noted that 3.1 percent have IQs below 70, usually considered the mentally retarded category. This column may be used to indicate the percentile rank (PR) of certain IQs. Thus an IQ of 119 has a PR of 89, since 89.4 percent of IQs are 119 or below; an IQ of 79 has a PR of 10.6, or 11. See DEVIATION IQ, MENTAL AGE.

Table 1. Normal Distribution of IQs with Mean of 100 and Standard Deviation of 16

(1) IQ Range	(2) Percent of Persons		(3) Cumulative Percent
140 and above	0.6		100.6
130-139	2.5		99.4
120-129	7.5		96.9
110-119	16.0		89.4
100-109	23.4	46.8	73.4
90- 99	23.4		50.0
80- 89	16.0		26.6
70- 79	7.5		10.6
60- 69	2.5		3.1
Below 60	0.6		0.6
Total	100.0		

internal consistency. Degree of relationship among the items of a test; consistency in content sampling. See SPLIT-HALF RELIABILITY.

interpolation. In general, any process of estimating intermediate values between two known points. As applied to test norms, it refers to the procedure used in assigning interpretive values (e.g., grade equivalents) to scores between the successive average scores actually obtained in the standardization process. Also, in reading norm tables it is necessary at times to interpolate to obtain a norm value for a score between two scores given in the table; e.g., in the table shown here, a percentile rank of 83 (from 81 + ⅓ of 6) would be assigned, by *interpolation*, to a score of 46; a score of 50 would correspond to a percentile rank of 94 (obtained as 87 + ⅗ of 10).

Score	Percentile Rank
51	97
48	87
45	81

inventory. A questionnaire or check list, usually in the form of a self-report, designed to elicit non-intellective information about an individual. Not tests in the usual sense, inventories are most often concerned with personality traits, interests, attitudes, problems, motivation, etc. See PERSONALITY TEST.

inventory test. An achievement test that attempts to cover rather thoroughly some relatively small unit of specific instruction or training. An inventory test, as the name suggests, is in the nature of a "stock-taking" of an individual's knowledge or skill, and is often administered prior to instruction.

item. A single question or exercise in a test.

item analysis. The process of evaluating single test items in respect to certain characteristics. It usually involves determining the difficulty value and the discriminating power of the item, and often its correlation with some external criterion.

Kuder-Richardson formula(s). Formulas for estimating the reliability of a test that are based on *inter-item consistency* and require only a single administration of the test. The one most used, formula 20, requires information based on the number of items in the test, the standard deviation of the total score, and the proportion of examinees passing each item. The Kuder-Richardson formulas are not appropriate for use with speeded tests.

mastery test. A test designed to determine whether a pupil has mastered a given unit of instruction or a single knowledge or skill; a test giving information on *what* a pupil knows, rather than on how his performance relates to that of some norm-reference group. Such tests are used in computer-assisted instruction, where their results are referred to as content- or criterion-referenced information.

mean (M). See ARITHMETIC MEAN.

median (Md). The middle score in a distribution or set of ranked scores; the point (score) that divides the group into two equal parts; the 50th percentile. Half of the scores are below the median and half above it, except when the median itself is one of the obtained scores.

mental age (MA). The age for which a given score on a mental ability test is average or normal. If the average score made by an unselected group of children 6 years, 10 months of age is 55, then a child making a score of 55 is said to have a mental age of 6–10. Since the mental age unit shrinks with increasing (chronological) age, MAs do not have a uniform interpretation throughout all ages. They are therefore most appropriately used at the early age levels where mental growth is relatively rapid.

modal-age norms. Achievement test norms that are based on the performance of pupils of normal age for their respective grades. Norms derived from such age restricted groups are free from the distorting influence of the scores of underage and overage pupils.

mode. The score or value that occurs most frequently in a distribution.

multiple-choice item. A test item in which the examinee's task is to choose the correct or best answer from several given answers or options.

N. The symbol commonly used to represent the number of cases in a group.

non-language test. See NON-VERBAL TEST.

non-verbal test. A test that does not require the use of words in the item or in the response to it. (Oral directions may be included in the formulation of the task.) A test cannot, however, be classified as non-verbal simply because it does not require reading on the part of the examinee. The use of non-verbal tasks cannot completely eliminate the effect of culture.

norm line. A smooth curve drawn to best fit (1) the plotted mean or median scores of successive age or grade groups, or (2) the successive percentile points for a single group.

normal distribution. A distribution of scores or measures that in graphic form has a distinctive bell-shaped appearance. Figures 1 and 2 show graphs of such a distribution, known as a *normal, normal probability,* or *Gaussian* curve. (Difference in shape is due to the different variability of the two distributions.) In such a normal distribution, scores or measures are distributed symmetrically about the mean, with as many cases up to various distances above the mean as down to equal distances below it. Cases are concentrated near the mean and decrease in frequency, according to a precise mathematical equation, the farther one departs from the mean. *Mean* and *median* are identical. The assumption that mental and psychological characteristics are distributed normally has been very useful in test development work.

norms. Statistics that supply a frame of reference by which meaning may be given to obtained test scores. Norms are based upon the actual performance of pupils of various grades or ages in the standardization group for the test. Since they represent average or typical performance, they should not be regarded as standards or as universally desirable levels of attainment. The most common types of norms are deviation IQ, percentile rank, grade equivalent, and stanine. Reference groups are usually those of specified age or grade.

objective test. A test made up of items for which correct responses may be set up in advance; scores are unaffected by the opinion or judgment of the scorer. Objective keys provide for scoring by clerks or by machine. Such a test is contrasted with a "subjective" test, such as the usual essay examination, to which different persons may assign different scores, ratings, or grades.

omnibus test. A test (1) in which items measuring a variety of mental operations are all combined into a single sequence rather than being grouped together by type of operation, and (2) from which only a single score is derived, rather than separate scores for each operation or function. Omnibus tests make for simplicity of administration, since one set of directions and one overall time limit usually suffice. The Elementary, Intermediate, and Advanced tests in the *Otis-Lennon Mental Ability Test* series are omnibus-type tests, as contrasted with the *Kuhlmann-Anderson Measure of Academic Potential,* in which the items measuring similar operations occur together, each with its own set of directions. In a *spiral-omnibus* test, the easiest items of each type are presented first, followed by the same succession of item types at a higher difficulty level, and so on in a rising spiral.

percentile (P). A point (score) in a distribution at or below which fall the percent of cases indicated by the percentile. Thus a score coinciding with the 35th percentile (P_{35}) is regarded as equaling or surpassing that of 35 percent of the persons in the group, and such that 65 percent of the performances exceed this score. "Percentile" has nothing to do with the percent of correct answers an examinee makes on a test.

percentile band. An interpretation of a test score which takes account of the measurement error that is involved. The range of such bands, most useful in portraying significant differences in battery profiles, is usually from one standard error of measurement below the obtained score to one standard error of measurement above it.

percentile rank (PR). The expression of an obtained test score in terms of its position within a group of 100 scores; the percentile rank of a score is the percent of scores equal to or lower than the given score in its own or in some external reference group.

performance test. A test involving some motor or manual response on the examinee's part, generally a manipulation of concrete equipment or materials. Usually *not* a paper-and-pencil test.

(1) A "performance" test of mental ability is one in which the role of language is excluded or minimized, and ability is assessed by what the examinee *does* rather than by what he says (or writes). Mazes, form boards, picture completion, and other types of items may be used. Examples include certain *Stanford-Binet* tasks, the Performance Scale of *Wechsler Intelligence Scale for Children, Arthur Point Scale of Performance Tests, Raven's Progressive Matrices.*

(2) "Performance" tests include measures of mechanical or manipulative ability where the task itself coincides with the objective of the measurement, as in the *Bennett Hand-Tool Dexterity Test.*

(3) The term "performance" is also used to denote a test that is actually a *work-sample;* in this sense it may include paper-and-pencil tests, as, for example, a test in bookkeeping, in shorthand, or in proofreading, where no materials other than paper and pencil may be required, and where the test response is identical with the behavior about which information is desired. *SRA Typing Skills* is such a test.

The use of the term "performance" to describe a type of test is not very precise and there are certain "gray areas." Perhaps one should think of "performance" tests as those on which the obtained differences among individuals may *not* be ascribed to differences in ability to use verbal symbols.

personality test. A test intended to measure one or more of the non-intellective aspects of an individual's mental or psychological make-up; an instrument designed to obtain information on the affective characteristics of an individual—emotional, motivational, attitudinal, etc.—as distinguished from his abilities. Personality tests include (1) the so-called *personality* and *adjustment inventories* (e.g., *Bernreuter Personality Inventory, Bell Adjustment Inventory, Edwards Personal Preference Schedule)* which seek to measure a person's status

[**personality test,** continued.]
on such traits as dominance, sociability, introversion, etc., by means of self-descriptive responses to a series of questions; (2) *rating scales* which call for rating, by one's self or another, the extent to which a subject possesses certain traits; and (3) *opinion or attitude inventories* (e.g., *Allport-Vernon-Lindzey Study of Values, Minnesota Teacher Attitude Inventory*). Some writers also classify interest, problem, and belief inventories as personality tests (e.g., *Kuder Preference Record, Mooney Problem Check List*). See PROJECTIVE TECHNIQUE.

power test. A test intended to measure level of performance unaffected by speed of response; hence one in which there is either no time limit or a very generous one. Items are usually arranged in order of increasing difficulty.

practice effect. The influence of previous experience with a test on a later administration of the same or a similar test; usually an increased familiarity with the directions, kinds of questions, etc. Practice effect is greatest when the interval between testings is short, when the content of the two tests is identical or very similar, and when the initial test-taking represents a relatively novel experience for the subjects.

predictive validity. See VALIDITY (2).

product-moment coefficient (r). Also known as the Pearson r. See COEFFICIENT OF CORRELATION.

profile. A graphic representation of the results on several tests, for either an individual or a group, when the results have been expressed in some uniform or comparable terms (standard scores, percentile ranks, grade equivalents, etc.). The profile method of presentation permits identification of areas of strength or weakness.

prognosis (prognostic) test. A test used to predict future success in a specific subject or field, as the *Pimsleur Language Aptitude Battery*.

projective technique (projective method). A method of personality study in which the subject responds as he chooses to a series of ambiguous stimuli such as ink blots, pictures, unfinished sentences, etc. It is assumed that under this free-response condition the subject "projects" manifestations of personality characteristics and organization that can, by suitable methods, be scored and interpreted to yield a description of his basic personality structure. The *Rorschach* (ink blot) *Technique*, the *Murray Thematic Apperception Test* and the *Machover Draw-a-Person Test* are commonly used projective methods.

quartile. One of three points that divide the cases in a distribution into four equal groups. The lower quartile (Q_1), or 25th percentile, sets off the lowest fourth of the group; the middle quartile (Q_2) is the same as the 50th percentile, or median, and divides the second fourth of cases from the third; and the third quartile (Q_3), or 75th percentile, sets off the top fourth.

r. See COEFFICIENT OF CORRELATION.

random sample. A sample of the members of some total population drawn in such a way that every member of the population has an equal chance of being included – that is, in a way that precludes the operation of bias or "selection." The purpose in using a sample free of bias is, of course, the requirement that the cases used be representative of the total population if findings for the sample are to be generalized to that population. In a *stratified* random sample, the drawing of cases is controlled in such a way that those chosen are "representative" also of specified subgroups of the total population. See REPRESENTATIVE SAMPLE.

range. For some specified group, the difference between the highest and the lowest obtained score on a test; thus a very rough measure of spread or variability, since it is based upon only two extreme scores. Range is also used in reference to the possible spread of measurement a test provides, which in most instances is the number of items in the test.

raw score. The first quantitative result obtained in scoring a test. Usually the number of right answers, number right minus some fraction of number wrong, time required for performance, number of errors, or similar direct, unconverted, uninterpreted measure.

readiness test. A test that measures the extent to which an individual has achieved a degree of maturity or acquired certain skills or information needed for successfully undertaking some new learning activity. Thus a *reading readiness* test indicates whether a child has reached a developmental stage where he may profitably begin formal reading instruction. *Readiness* tests are classified as *prognostic* tests.

recall item. A type of item that requires the examinee to supply the correct answer from his own memory or recollection, as contrasted with a *recognition item*, in which he need only identify the correct answer.

Columbus discovered America in the year _____

is a *recall* (or *completion*) item. See RECOGNITION ITEM.

recognition item. An item which requires the examinee to recognize or select the correct answer from among two or more given answers (options).

Columbus discovered America in
(a) *1425* (b) *1492* (c) *1520* (d) *1546*

is a *recognition* item.

regression effect. Tendency of a predicted score to be nearer to the mean of its distribution than the score from which it is predicted is to its mean. Because of the effects of regression, students making extremely high or extremely low scores on a test tend to make less extreme scores, i.e., closer to the mean, on a second administration of the same test or on some predicted measure.

reliability. The extent to which a test is consistent in measuring whatever it does measure; dependability, stability, trustworthiness, relative freedom from errors of measurement. Reliability is usually expressed by some form of *reliability coefficient* or by the *standard error of measurement* derived from it.

reliability coefficient. The coefficient of correlation between two forms of a test, between scores on two administrations of the same test, or between halves of a test, properly corrected. The three measure somewhat different aspects of reliability, but all are properly spoken of as reliability coefficients. See ALTERNATE-FORM RELIABILITY, SPLIT-HALF RELIABILITY COEFFICIENT, TEST-RETEST RELIABILITY COEFFICIENT, KUDER-RICHARDSON FORMULA(S).

representative sample. A sample that corresponds to or matches the population of which it is a sample with respect to characteristics important for the purposes under investigation. In an achievement test norm sample, such significant aspects might be the proportion of cases of each sex, from various types of schools, different geographical areas, the several socioeconomic levels, etc.

scholastic aptitude. See ACADEMIC APTITUDE.

skewed distribution. A distribution that departs from symmetry or balance around the mean, i.e., from normality. Scores pile up at one end and trail off at the other.

Spearman-Brown formula. A formula giving the relationship between the reliability of a test and its length. The formula permits estimation of the reliability of a test lengthened or shortened by any multiple, from the known reliability of a given test. Its most common application is the estimation of reliability of an entire test from the correlation between its two halves. See SPLIT-HALF RELIABILITY COEFFICIENT.

split-half reliability coefficient. A coefficient of reliability obtained by correlating scores on one half of a test with scores on the other half, and applying the Spearman-Brown formula to adjust for the doubled length of the total test. Generally, but not necessarily, the two halves consist of the odd-numbered and the even-numbered items. Split-half reliability coefficients are sometimes referred to as measures of the *internal consistency* of a test; they involve content sampling only, not stability over time. This type of reliability coefficient is inappropriate for tests in which speed is an important component.

standard deviation (S.D.). A measure of the variability or dispersion of a distribution of scores. The more the scores cluster around the mean, the smaller the standard deviation. For a normal distribution, approximately two thirds (68.3 percent) of the scores are within the range from one S.D. below the mean to one S.D. above the mean. Computation of the S.D. is based upon the square of the deviation of each score from the mean. The S.D. is sometimes called "sigma" and is represented by the symbol σ. (See Figure 1.)

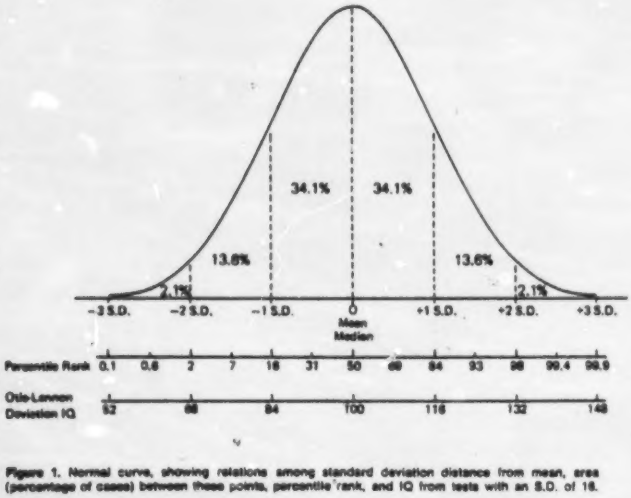

Figure 1. Normal curve, showing relations among standard deviation distance from mean, area (percentage of cases) between these points, percentile rank, and IQ from tests with an S.D. of 16.

standard error (S.E.). A statistic providing an estimate of the possible magnitude of "error" present in some obtained measure, whether (1) an *individual* score or (2) some *group* measure, as a mean or a correlation coefficient.

(1) standard error of measurement (S.E. Meas.): As applied to a single obtained score, the amount by which the score may differ from the hypothetical true score due to errors of measurement. The larger the S.E. Meas., the less reliable the score. The S.E. Meas. is an amount such that in about two-thirds of the cases the obtained score would not differ by more than one S.E. Meas. from the true score. (Theoretically, then, it can be said that the chances are 2:1 that the actual score is within a band extending from *true score minus 1 S.E. Meas.* to *true score plus 1 S.E. Meas.*; but since the true score can never be known, actual practice must reverse the true-obtained relation for an interpretation.) Other probabilities are noted under (2) below. See TRUE SCORE.

(2) standard error: When applied to group averages, standard deviations, correlation coefficients, etc., the S.E. provides an estimate of the "error" which may be involved. The group's size and the S.D. are the factors on which these standard errors are based. The same probability interpretation as for S.E. Meas. is made for the S.E.s of group measures, i.e., 2:1 (2 out of 3) for the 1 S.E. range, 19:1 (95 out of 100) for a 2 S.E. range, 99:1 (99 out of 100) for a 2.6 S.E. range.

standard score. A general term referring to any of a variety of "transformed" scores, in terms of which raw scores may be expressed for reasons of convenience, comparability, ease of interpretation, etc. The simplest type of standard score, known as a z-score, is an expression of the *deviation* of a score from the mean score of the group *in relation to* the standard deviation of the scores of the group. Thus:

$$\text{standard score (Z)} = \frac{\text{raw score (X)} - \text{mean (M)}}{\text{standard deviation (S.D.)}}$$

Adjustments may be made in this ratio so that a system of standard scores having any desired mean and standard deviation may be set up. The use of such standard scores does not affect the relative standing of the individuals in the group or change the shape of the original distribution. T-scores have a M of 50 and an S.D. of 10. Deviation IQs are standard scores with a M of 100 and some chosen S.D., most often 16; thus a raw score that is 1 S.D. above the M of its distribution would convert to a standard score (deviation IQ) of $100 + 16 = 116$. (See Figure 1.)

Standard scores are useful in expressing the raw scores of two forms of a test in comparable terms in instances where tryouts have shown that the two forms are not identical in difficulty; also, successive levels of a test may be linked to form a continuous standard-score scale, making across-battery comparisons possible.

standardized test (standard test). A test designed to provide a systematic sample of individual performance, administered according to prescribed directions, scored in conformance with definite rules, and interpreted in reference to certain normative information. Some would further restrict the usage of the term "standardized" to those tests for which the items have been chosen on the basis of experimental evaluation, and for which data on reliability and validity are provided. Others would add "commercially published" and/or "for general use."

stanine. One of the steps in a nine-point scale of standard scores. The stanine (short for *standard-nine)* scale has values from 1 to 9, with a mean of 5 and a standard deviation of 2. Each stanine (except 1 and 9) is ½ S.D. in width, with the middle (average) stanine of 5 extending from ¼ S.D. below to ¼ S.D. above the mean. (See Figure 2.)

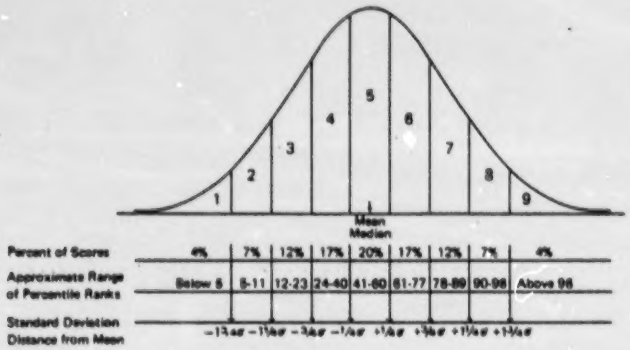

Figure 2. Stanines and the normal curve. Each stanine (except 1 and 9) is one half S.D. in width.

survey test. A test that measures general achievement in a given area, usually with the connotation that the test is intended to assess group status, rather than to yield precise measures of individual performance.

t. A critical ratio expressing the relationship of some measure (mean, correlation coefficient, difference, etc.) to its standard error. The size of this ratio is an indication of the significance of the measure. If *t* is as large as 1.96, significance at the .05 level is indicated; if as large as 2.58, at the .01 level. These levels indicate 95 or 99 chances out of 100, respectively.

taxonomy. An embodiment of the principles of classification; a survey, usually in outline form, such as a presentation of the objectives of education.

test-retest reliability coefficient. A type of reliability coefficient obtained by administering the same test a second time, after a short interval, and correlating the two sets of scores. "Same test" was originally understood to mean identical content, i.e., the same form; currently, however, the term "test-retest" is also used to describe the administration of different forms of the same test, in which case this reliability coefficient becomes the same as the alternate-form coefficient. In either case (1) fluctuations over time and in testing situation, and (2) any effect of the first test upon the second are involved. When the time interval between the two testings is considerable, as several months, a test-retest reliability coefficient reflects not only the consistency of measurement provided by the test, but also the stability of the examinee trait being measured.

true score. A score entirely free of error; hence, a hypothetical value that can never be obtained by testing, which always involves some measurement error. A "true" score may be thought of as the average score from an infinite number of measurements from the same or exactly equivalent tests, assuming no practice effect or change in the examinee during the testings. The standard deviation of this infinite number of "samplings" is known as the *standard error of measurement.*

validity. The extent to which a test does the job for which it is used. This definition is more satisfactory than the traditional "extent to which a test measures what it is supposed to measure," since the validity of a test is always specific to the purposes for which the test is used. The term validity, then, has different connotations for various types of tests and, thus, a different kind of validity evidence is appropriate for each.

(1) content, curricular validity. For achievement tests, validity is the extent to which the *content* of the test represents a balanced and adequate sampling of the outcomes (knowledge, skills, etc.) of the course or instructional program it is intended to cover. It is best evidenced by a comparison of the test content with courses of study, instructional materials, and statements of educational goals; and often by analysis of the processes required in making correct responses to the items. *Face validity,* referring to an observation of what a test appears to measure, is a non-technical type of evidence; apparent relevancy is, however, quite desirable.

(2) criterion-related validity. The extent to which scores on the test are in agreement with *(concurrent validity)* or predict *(predictive validity)* some given criterion measure. Predictive validity refers to the accuracy with which an aptitude, prognostic, or readiness test indicates future learning success in some area, as evidenced by correlations between scores on the test and future criterion measures of such success (e.g., the relation of score on an academic aptitude test administered in high school to grade point average over four years of college). In concurrent validity, no significant time interval elapses between administration of the test being validated and of the criterion measure. Such validity might be evidenced by *concurrent* measures of academic ability and of achievement, by the relation of a new test to one generally accepted as or known to be valid, or by the correlation between scores on a test and criteria measures which are valid but are less objective and more time-consuming to obtain than a test score would be.

(3) construct validity. The extent to which a test measures some relatively abstract psychological trait or construct; applicable in evaluating the validity of tests that have been constructed on the basis of an analysis (often factor analysis) of the nature of the trait and its manifestations. Tests of personality, verbal ability, mechanical aptitude, critical thinking, etc., are validated in terms of their construct and the relation of their scores to pertinent external data.

variability. The spread or dispersion of test scores, best indicated by their standard deviation.

variance. For a distribution, the average of the squared deviations from the mean; thus the square of the standard deviation.

TEST SERVICE NOTEBOOKS are issued from time to time as a professional service of The Psychological Corporation. Inquiries, comments, or requests for additional copies may be addressed to the office nearest you. Write: Advisory Services, The Psychological Corporation, New York, NY 10017 • Chicago, Il 60648 • San Francisco, CA 94109 • Atlanta, GA 30309 • Dallas, TX 75235

Appendix B

Summary of Common Test Scores

SCORES FREQUENTLY ASSOCIATED WITH NORM REFERENCED TESTS

	DEFINITION	MAJOR ADVANTAGES	MAJOR DISADVANTAGES
PERCENTILE RANK	The percentile rank establishes a student's standing relative to a norm group in terms of the percentage of students who scored at or below his or her raw score. For example, a student who scored at the 98th percentile achieved a raw score which was higher than the raw scores of 98 percent of the norm group who took the same test under the same conditions.	1. Percentiles show the relative standing of individuals compared to a normative group. 2. They are familiar to most public school personnel, though probably not the general public. 3. Percentiles are relatively easily explained.	1. Percentiles are frequently confused with the percent of the total number of test items answered correctly. 2. Since the percentile scale does not have equal units of measurement, percentiles should not be used in the computation of group statistics.
GRADE EQUIVALENT SCORE	The grade equivalent score indicates the performance of a student on a particular test relative to the median performance of students at a given grade level and month; e.g., a fifth grader who receives a grade equivalent score of 8.2 on a reading test achieved the same raw score performance as the typical eighth grader in the second month of eighth grade would be expected to achieve on the same fifth grade test.	1. It appears easy to communicate the standing of an individual student relative to a grade level (most people believe they understand what is meant by grade equivalent scores).	1. Grade equivalents are easily misunderstood and misinterpreted. 2. Achievement expressed in grade equivalent score units cannot be meaningfully compared with each other in several instances. a. Grade equivalent scores cannot be meaningfully compared for the same student (or group of students) over time. b. Grade equivalent scores cannot be meaningfully compared for the same student (or group of students) across subject matter areas. c. Grade equivalent scores cannot be meaningfully compared for the same student (or group of students) across different tests. 3. Many grade equivalent scores are statistical projections (interpolations or extrapolations). In the later grades it is not uncommon to find grade equivalent scores of two or three grade levels above or below the student's actual grade level, but these scores are of doubtful accuracy. 4. The grade equivalent scale is not composed of equal sized units. Having equal sized units implies that the underlying difference between any two scores is the same throughout the scale.

SCORES FREQUENTLY ASSOCIATED WITH NORM REFERENCED TESTS

	DEFINITION	MAJOR ADVANTAGES	MAJOR DISADVANTAGES
STANDARD SCORE	Standard scores are derived from raw scores, but express the results of a test on the same numerical scale regardless of grade level, subject area or test employed.	1. Since the mean and standard deviation of the standard score scales are pre-specified, a student's standard score immediately communicates two important facts about his or her performance on that test: a. Whether the student's score is above or below the mean. b. How far above or below the mean, in standard deviation units, his or her performance is. 2. The constant numerical scale of standard scores facilitates comparisons: a. Across students taking the same test. b. Across subject matter areas for the same student. 3. Standard scores are derived in a way that maintains the equal interval property in their units which is absent in percentile and grade equivalent scores. Therefore, summary statistics may be meaningfully interpreted when calculated on standard scores.	1. The most useful interpretation of standard scores requires some knowledge of statistics (i.e., mean and standard deviation) and hence may not be appropriate for audiences who have not been exposed to these concepts (e.g., parents, the news media). 2. Given the variety of standard scores available, there may be potential confusion in expressing the same test performance with so many different numerical values. 3. The conversion of raw scores to standard scores may either maintain the shape of the distribution observed, or may transform the distribution to another, more interpretively convenient shape (e.g., the normal distribution); and the procedures employed in specifying the conversion process may not be immediately obvious.
NORMAL CURVE EQUIVALENTS	A standard score system having 99 equal intervals. The average corresponds to the 50th centile; the 1st & 99th NCEs correspond to the 1st & 99th centiles. Range: generally 1-99 but can be higher and lower.	1. Same as standard score systems. 2. Permit aggregation of data from a wide variety of tests.	1. They are relatively new. 2. They depend upon standard scores or percentiles. 3. Not all test publishers use them.

SCORES FREQUENTLY ASSOCIATED WITH NORM REFERENCED TESTS

	DEFINITION	MAJOR ADVANTAGES	MAJOR DISADVANTAGES
EXPANDED SCALE SCORE	Expanded scale scores are a type of standard score whose scale is designed to extend across grade levels and whose mean increases progressively as the grade level increases.	1. Expanded scores facilitate longitudinal comparisons of an individual across grade levels.	1. Different test publishers use different terms to refer to their expanded scale scores (e.g., growth scale values, achievement development scale scores, standard score, scale score) and this may be confusing when considering results from different tests.
		2. Expanded scale scores provide the vehicle for expressing a performance obtained at one grade level to the norm group of another. This is useful when the appropriate level of a test to be administered to a student is judged to be other than that of his or her grade level (i.e., functional level testing).	2. Different tests use different ranges, and standard deviations in deriving their expanded scale scores. Thus, results from different tests expressed in expanded scale score units cannot be readily compared.
		3. Since they were designed as equal interval, their scores may be mathematically manipulated (e.g., averaged).	3. The statistical properties of expanded scale scores are often not as uniform as theoretically desired.
STANINE	Stanines are a standard score scale consisting of nine values with a mean of five and a standard deviation of two. If the distribution of scores is normal, each stanine includes a known proportion of the scores in the distribution.	1. As in all standard scores, stanines have the same meaning across different tests, different grade levels and different content areas.	1. Since some of the stanines encompass a wide range of scores, their use in reporting can be insensitive to differences between students' performance that are more apparent from the use of other test scores.
		2. Stanines consist of only nine possible scores and thus may be easier to communicate to audiences not familiar with measurement terminology. Verbal labels may be given to each stanine value to facilitate interpretation.	

SCORES FREQUENTLY ASSOCIATED WITH OBJECTIVE REFERENCED TESTS

	DEFINITION	MAJOR ADVANTAGES	MAJOR DISADVANTAGES
RAW SCORE	The number of items on a test or subtest answered correctly by the student.	1. Virtually no statistical or measurement expertise is needed to calculate raw scores. 2. Raw scores are the necessary first step in expressing test performance in any of a number of other ways (e.g., standard scores, percentiles.)	1. By themselves, raw scores offer no indication as to how a student who has mastered the skills represented on the test "should" perform (i.e., criterion referenced) or how other students at the same grade level have performed (i.e., norm referenced.)
% ITEMS CORRECT	The proportion of the total number of items answered correctly by the student.	1. Very little statistical or measurement expertise is required to understand this expression of test performance. 2. If the content area is sufficiently represented by the items on the test, the percent correct provides an expression of the proportion of the subject matter mastered by the student.	1. No notion of test difficulty or expected performance is contained in this score. Unless accompanied by a standard for mastery or information as to how a student's peers have performed in the test, misinterpretations may arise.
OBJECTIVE MASTERY SCORE	When a standard for mastery has been applied to a set of items for a specific objective, a student's performance in terms of that objective is expressed as having mastery or non-mastery of the objective.	1. The objective mastery score compares the student's performance on that objective to a judged standard of what he or she should know of the skills required to master it. This score can be very useful in diagnosing a student's specific strengths and weaknesses. 2. When the subject matter requires a successive accumulation of skills (e.g., elementary math), objective mastery scores may be extremely useful in monitoring the progress of students in specific skill areas.	1. Objective mastery scores are difficult to compare across different tests. Items designed to measure the same objective may differ in difficulty or have different standards for mastery on different tests. 2. If a purpose in testing is to differentiate among students, objective mastery scores do not present a very useful index. Different raw scores above or below the mastery level are viewed as the same--either mastery or non-mastery.